DOCUMENT RESUME

ED 387 287 RC 020 232

AUTHOR Valencia, Richard R.; Aburto, Sofia
TITLE The Uses and Abuses of Educational Testing: Chicanos as a Case in Point. Chapter 8.
PUB DATE 93
NOTE 51p.; In: Chicano School Failure and Success: Research and Policy Agendas for the 1990s; see RC 020 224.
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Culture Fair Tests; Educational Research; *Educational Testing; Elementary Secondary Education; Evaluation Research; Higher Education; Intelligence Tests; *Mexican American Education; Mexican Americans; Minimum Competency Testing; Minority Group Children; Standardized Tests; *Student Evaluation; Teacher Competency Testing; *Test Bias; *Test Use
IDENTIFIERS *Chicanos; Hispanic American Students; *Nondiscriminatory Assessment

ABSTRACT

A persistent problem in U.S. educational research has been how to explain the continuing low performance on standardized tests by certain racial and ethnic minority-group students, such as Chicanos. This chapter identifies abusive practices stemming from standardized testing that help to shape school failure among Chicano students, and discusses proactive research and policy strategies to enhance Chicano school success. Following a brief overview of the functions of educational testing, test abuse with respect to Chicano students is analyzed for intelligence tests and competency based tests. The section on intelligence testing covers history, early efforts in the 1960s and 1970s to provide nondiscriminatory assessment for Chicano students, test bias research and Chicanos, the responsibilities of test publishers and school psychologists in helping to promote nondiscriminatory assessment, and the need to link nondiscriminatory assessment with nondiscriminatory schooling. The section on competency testing covers minimum competency tests used as standards for high school graduation and teacher competency tests, and focuses on concerns about the psychometric integrity of tests and the reliance on tests as the sole or primary source of data for educational decision making. Impacts on minority group students and on the supply of Chicano and bilingual teachers are examined. Along with abusive testing practices, the notion of limited "educability" also is instrumental in creating barriers to educational opportunities for Chicanos. Eight research and policy oriented ideas are offered to improve educational testing and promote Chicano school success. Contains 145 references. (SV)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 387 287

PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

E. STREISAND
(FALMER PRESS)

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

RC 020232

BEST COPY AVAILABLE

Chapter 8

The Uses and Abuses of Educational Testing: Chicanos as a Case in Point

Richard R. Valencia and Sofia Aburto

One of the most persistent problems in educational research in the United States has been how to explain the continuing low performance on standardized tests by certain racial/ethnic minority-group students, such as Chicanos. Second, the potential uses and abuses of educational tests with Chicanos and other minorities have generated tremendous controversy over the years. These two major testing issues — correlates and consequences of testing performance *vis-à-vis* Chicano students — conform one of the most profound and controversial debates in the annals of education, a debate that has spilled beyond the confines of the academic community. The media, public, courts, and legislative bodies have all entered the fray in one form or another. In short, these testing issues are historically rooted, controversial and — by their pervasiveness — important within and outside the institution of education.

Although the correlates associated with standardized test performance of Chicano students are important to examine in reaching some understanding of Chicano school failure and success, they will only be lightly touched upon in this chapter (see Laosa and Henderson, this volume, for a discussion of socialization and competence aspects of cognitive performance). Our goal here is to identify and discuss a number of abusive practices stemming from standardized testing that we believe help shape school failure among Chicano students.[1] We will not, however, dwell entirely on the negative. Given the spirit and charge of the present book, our focus will also be on the identification of proactive ideas about educational testing, particularly in the form of research and policy strategies that are likely to enhance school success among Chicano students.

The discussion begins with a brief overview of the functions of testing. This follows with the chapter's core — an analysis of test abuse with respect to Chicano students. To do this, we employ a 'test typology' format. That is, our focus on abusive practices is placed in the context of the two following types of tests: 'intelligence' and 'competency' based.[2] The section on intelligence testing covers a brief history, the early period of nondiscriminatory assessment, test bias research and Chicano students, the responsibility of test publishers and school psychologists in helping to promote nondiscriminatory assessment, and the need to link nondiscriminatory assessment with nondiscriminatory schooling. The

section on competency testing examines the various types of competency tests (e.g., minimum competency tests; teacher competency examinations), and discusses their impact on Chicanos. Following this is a discussion of the notion of 'educability'. The chapter closes with a presentation of eight research and policy ideas about improving educational testing in order to help promote Chicano school success.

Functions of Testing

Before any analysis of test abuse is undertaken, a logical question to ask is, what functions do educational tests serve? There are a number of frameworks that have been advanced to address this concern (e.g., Cronbach, 1984; Resnick, 1979; Salvia and Ysseldyke, 1988; Thorndike and Hagen, 1977).[3] The analysis we find particularly useful for the issues discussed here is the framework advanced by Resnick, who presents a lucid discussion on current test use in the schools.[4] Resnick identifies three broad test functions: 1) the management of instruction, 2) public accountability, and 3) the legitimization of the schooling process.

Management of Instruction

This is the rubric for several purposes of testing — the sorting, monitoring, and grading functions. In the sorting function, tests are given before the instructional process begins. Here, educational tests serve as mechanisms to assist in the assignment of students to special education and, as Resnick (1979) notes, for 'tracking' in the educational mainstream. With respect to the monitoring function, tests are administered during the course of the instructional process and provide information that can be used to make curriculum adjustments so as to improve student achievement. The third purpose of testing within the management of instruction area is the grading function. Here, tests are given at the end of the instructional process and serve as sources of evaluating a student's academic performance.

In the case of the Chicano, as well as other minority students (e.g., Black), the sorting function has created the most controversy. For example, the issue of overrepresentation of racial/ethnic minority children in classes for the educable mentally retarded (EMR) was particularly explosive during the 1970s (Henderson and Valencia, 1985; see also Rueda, this volume). With respect to Chicano students, the EMR misclassification of many children from this ethnic population during the 1970s was likely the outgrowth of a long taproot. There is historical evidence from the 1920s that IQ tests were routinely used as sorting instruments to place large numbers of Chicano pupils in classes for the 'mentally defective' (Gonzalez, 1974a). More on the abuses of intelligence testing will be presented in the section, 'Intelligence Testing and Chicano Students: A Brief History'.

Public Accountability

The general notion of accountability in education is that public schools should be held accountable to the public (the logic being that the public financially supports

the schools). Milliken (1970) describes this idea as a collective sense '... that people are increasingly demanding to know how their children are learning, what they are learning, and why they are being taught whatever they are being taught' (p. 17). Norm-referenced achievement and aptitude tests are typically used to meet the public's demands for accountability. In a later section, we will discuss the abuses of competency-based testing *vis-à-vis* Chicanos. (Competency testing, broadly conceptualized, refers to the testing of examinees for the acquisition of basic skills. Often such tests are used as gatekeepers for determining grade-to-grade promotion, graduation, entry to pre-professional training programs, and so on.)

Legitimization of the Schooling Process

This is the third broad function of educational testing that Resnick (1979) discusses in her framework. In that the legitimization function of testing is linked to our broader analysis of test abuse, an expanded discussion of this function is necessary at this time. We will take intelligence testing as a case in point.

Each of our nation's 16,000 public school districts maintains and supports a program of standardized testing. The testing movement in the United States, beginning in the second decade of the twentieth century, is deeply rooted in our desire for efficiency, our ideas of equality, and our need to have national standards (Resnick, 1981). The notions and values of 'rational management', 'scientific management', and 'efficiency reform' as applied to public education at the turn of the century stem from the larger influence of the business ideology and the application of modern business methods during the Progressive period from 1890 to 1920 (Callahan, 1962). These values of rationality and efficiency were initially reactions to corruption and inefficiency in government, but because of huge problems in public schools (e.g., high rates of dropouts, overcrowding), '... schools became a central target for the efficiency reformers in the decade before World War I' (Resnick, 1981, p. 625). The scientific management ethos and the use of intelligence testing in the schools took on massive proportions as seen in the creation of numerous bureaus of research and measurement from 1912 to 1922 in urban school systems (Resnick, 1981).

In a lucid account of how the intelligence testing movement transformed administrative policies in public schools, Tyack (1974) notes that one survey in 1926 reported thirty-seven out of forty cities with populations of 100,000 or more were using intelligence tests for ability grouping in some or all elementary and secondary schools. By 1932, three-fourths of 150 large cities made curricular assignments of pupils by using the results of intelligence tests. Historical case examinations of selected cities (e.g., Gonzalez's (1974a), study of the testing program of Los Angeles City School District and its effect on Chicano students from 1920 to 1930) illustrate the inner workings and powerful influences of early intelligence testing and research departments on the educational bureaucracies of urban school systems. As Resnick (1981) concludes on this topic:

> The present use of testing to support decisions about ability and curriculum grouping affirms traditions of practice that have been in existence for more than 60 years ... The American use of tests reflects our

> culture's interest in qualified and 'objective judgments, part of the rational management ethos'. (p. 626)

In light of this marriage of scientific management and the intelligence testing movement, it becomes clearer why Resnick (1979) believes that intelligence testing has served a 'legitimization of the schooling process' function. The point is that in addition to having a practical function in schooling, intelligence tests also play symbolic roles through their aura of science and objectivity.[5] The implications here are important. If intelligence tests are indeed objective (in content and use of results), then our highly differentiated and tracked public school system is sanctioned. Bowles and Gintis (1976), writing from a neo-Marxist perspective, have taken the legitimization function of intelligence testing a bit further. In brief, these scholars offer the following argument: (a) evidence shows that test scores (IQ) are poor predictors of individual economic success; (b) the meritocratic mechanism — test scores — are assumed to be objective; (c) because economic success, however, cannot be accounted for by cognitive scores of students, then the technocractic-meritocratic ideology is largely symbolic and is used to legitimize economic inequality.

In summary, the notion that intelligence and other forms of educational testing perform a legitimization function is a powerful idea. Its utility as an analytical tool in understanding the abusive practices of education testing with respect to Chicanos will be discussed in the next section.

Abuses of Educational Testing

As Linn, Madaus and Pedulla (1982) underscore, 'When properly used, a test can be a valuable educational tool' (p. 1). The extent tests contribute to improving schooling for students depends on their psychometric integrity (i.e., reliability and validity) as well as their proper interpretation and use. As do Linn *et al.*, we also support test use along certain lines, but we are as strongly opposed to test misuse — and the resultant abusive consequences. Our primary criticisms of test misuse center on two concerns. First is the issue of administering tests that lack good, intrinsic quality — that is, the administration of unreliable tests and tests that have not been validated for specific uses (a subject we later discuss in some detail). Second, there is concern that tests are often used as the sole or major determinant in educational decision-making. A number of scholars and professional organizations have denounced this exclusive, or almost exclusive, reliance of a single test source in decision-making as an improper use of educational testing (e.g., International Reading Association, 1979; Linn *et al.*, 1982).

In this section, we will focus on test abuse with respect to Chicanos in the two broad areas of intelligence testing and competency testing. The latter will cover issues pertinent to minimum competency tests used as standards for high school graduation and teacher competency tests. Throughout the discussion, we will weave in the above concerns about the psychometric integrity of tests and the reliance on tests as the sole or primary data source in the making of educational decisions. Following this section there will be a discussion of the notion of 'educability', a concept we argue that is quite central in the analysis of educational testing and Chicano students.

Intelligence Testing

Intelligence Testing and Chicano Students: A Brief History

In the twentieth century, the construct of intelligence has been given more research attention by psychologists than any other dimension of individual differences (Peterson, 1982). Aside from the sheer volume of this research, more significant have been the long-lasting debates on the 'origins' of intelligence (especially racial/ethnic group differences) and the uses and abuses of intelligence testing results.

It is not the intent here to add further analysis, in any great detail, to the questions and research dealing with the correlates and possible influences on Chicano intellectual performance. Our discussion will be largely confined to an overview of the consequences of test abuse (for a much broader and sustained analysis of Chicano intellectual performance with respect to research, theory, and schooling implications, see Valencia, 1990). This is an attempt to shed some light on questions, such as, in general, what have been the psychological and social consequences of intelligence testing for Chicano students? Assuming that some of this impact has been negative, how has it contributed in part to Chicano school failure? To address these concerns, it is first necessary to look back in time, and then to place our eyes on the contemporary scene.

In that group-administered intelligence tests have been widely banned across the nation for nearly two decades, it is necessary to examine history to understand the foundations of oppressive test use. There is abundant evidence from scholarly work that the results of intelligence tests — particularly during the 1920s and 1930s — were used in racially discriminatory ways *vis-à-vis* Chicano and other racial/ethnic minority students (e.g., Blum, 1978; Gonzalez, 1974a, 1974b; Henderson and Valencia, 1985; Hendrick, 1977; Kamin, 1974; Valencia, 1990). During the heyday of testing in the 1920s and 1930s, practically all large cities in the United States had massive educational bureaucracies routinely administering group-based intelligence tests.

For example, Gonzalez (1974a, 1974b) found that the institutionalization of IQ testing, tracking, curriculum development, and counseling programs were used in ways that effectively stratified students along socioeconomic and racial/ethnic lines.[6] Given their typically and consistently low performance on intelligence tests,[7] Chicano students were often funnelled into slower tracks that frequently led to a low-grade 'vocational education' curriculum.[8] Furthermore, Gonzalez notes that Chicano children who scored below an IQ of 70 on standardized intelligence tests were referred to 'development centers' for the 'mentally retarded'.[9] In short, it appears that Chicano students in the Los Angeles public schools in the 1920s and 1930s routinely faced one of two equally unattractive educational paths — non-academic vocational education that emphasized low-level skills or dead-end special education. Offering a macrolevel analysis of the linkages between intelligence testing and consequences for Chicanos, Gonzalez (1974b) ties up matters this way:

> On the basis of IQ tests administered by guidance counselors, inordinate numbers of Mexican-American children were placed in coursework which prepared them for a variety of manual operations ... This movement was a reaction of the privileged classes to the rising numbers of the

> working classes. Schools were redefined in the era of monopoly capitalism to be instruments through which social order could be preserved and industrialization expanded. Thus American schools were not and still are not agents of change, but rather bolster the social stratifications and values of our society. In such an educational system, Mexican-Americans were not provided with opportunities to improve their lot but instead were subjected to a socialization process that reinforced the status quo and was opposed to social change. (p. 301)

In sum, historical research provides some evidence about the abusive practices of intelligence testing with respect to Chicano students. It would be incorrect to argue that the invidious misclassifications and channeling of Chicanos into unchallenging, low-status curricula depended *exclusively* on IQ tests. Yet, such tests did have a role — along with other elements (e.g., preconceived notions about Chicano children's educability; forced school segregation of Chicanos by the White community) — in helping shape inferior schooling for Chicanos. Whether intelligence testing in later years has negatively affected the 'quality of life' for Chicanos to the great degree some have claimed (e.g., Aguirre, 1979a, 1979b), is certainly a claim open for debate (a point we will return to later when we discuss 'educability').

Before moving on to a discussion of contemporary intelligence testing issues with respect to Chicano students, we need to provide a brief description for the period sandwiched between the heyday of IQ testing during the 1920s and 1930s and the group IQ testing ban of the 1970s. The period from the 1940s to the early 1960s represents a time in which intelligence testing did not receive much attention in either scholarly or popular writings (Haney, 1981). Part of this inattention, we think, is likely related to the notion that group-based intelligence testing after the 1930s became widely implemented in the nation's schools and took on a life of its own — a life relatively free of controversy.[10]

To some degree, the entrenchment and solidification of group-administered intelligence testing in the middle of the twentieth century had its curricular roots in the nascent period of testing. Fass (1980) notes that the purpose and direction of the intelligence testing movement during the early decades of the century were driven by a rapidly changing complex society and educational system in serious need of a socially '... powerful organizing principle' (p. 432). Urbanization, industrialization, and massive immigration combined to force the schools to address a critical issue of democratic schooling, which was, as Fass argues: '... how to educate the mass without losing sight of the individual' (p. 446). Stated in another manner, how could the United States educate a presumably intellectually diverse student population while sorting, selecting, and rewarding individual talent in a democratically and scientifically defensible manner? The answer, according to Fass, was the idea of IQ and intelligence testing, which coincidentally, were becoming available at a time when an organizing mechanism for selection was so much desired.

As time progressed, intelligence testing became widely used, serving as a sorting mechanism in the educational mainstream as well as the tributary of special education (see Rueda, this volume, for an analysis and critique of special education with respect to Chicano students). In the absence of hard data, it is not possible to claim unequivocally that group-based IQ tests during the 1940s, 50s,

and 1960s were used as unambiguous instruments to stratify students along lines of differentiated curricula. There is no denial that such testing played a role in ability grouping at the elementary level and in tracking at the secondary level. What is not known is how the day-to-day process of curriculum differentiation was influenced by IQ testing, particularly the *degree* to which teachers and counselors relied on test results to make placement and instructional decisions. What we can say, however, is that intelligence testing — along with other institutionalized mechanisms, such as school segregation — helped to limit the learning opportunities for Chicano students and thus contributed in shaping Chicano school failure.

Around 1960, public attention to testing peaked once again (Haney, 1981). Spurred by the launching of Sputnik in 1957 and the subsequent beginning of the 'space race', the identification of 'academically talented' youth became a national obsession. Large-scale testing proliferated, and soon after articles critical of intelligence and other testing (e.g., National Merit Scholarship) appeared in the popular literature.

In the late 1960s, one of the hottest debates of modern time related to IQ testing dealt with the issue of racial differences. Although the 'nature' v. 'nurture' debate over intelligence has occupied the annals of science for over a century (Blum, 1978), it was rekindled with new force by Jensen's (1969) controversial monograph on Black-White differences in intellectual ability. In a lengthy treatise, Jensen hypothesized that the lower intellectual performance of Black Americans was largely due to genetic influences. With a seeming momentum of its own, the nature v. nurture controversy keeps rolling along, as indicated by numerous works in the 1970s and even more recently (e.g., Dunn, 1987; Eysenck and Kamin, 1981; Flynn, 1980; Mercer, 1988). For example, Dunn recently wrote that Hispanic-Anglo differences in intellectual performance are largely due to genetic differences in intelligence. (For a critique of Dunn's position, see the entire issue of the *Hispanic Journal of Behavioral Sciences*, 1988, **10**, 3.)

Nondiscriminatory Assessment and Chicano Students: The Early Years

In 1964, the Society of Social Issues (Division 9 of the American Psychological Association) presented one of the first attempts in modern times to clarify discriminatory assessment issues (Deutsch, Fishman, Kogan, North and Whiteman, 1964). The authors called for greater sensitivity, responsibility, and goodwill on the part of those who test minority children. Concern about discriminatory assessment continued to some degree, but it was not until the early 1970s that the issues of cultural bias in intelligence tests and misclassification of minority children arrived on the national scene. Professional associations, litigation, and legislation were three major influences that helped define the issues and helped fashion more appropriate psychoeducational assessment and services for handicapped and racial/ethnic minority children (Henderson and Valencia, 1985; Oakland and Laosa, 1977; Reschly, 1980).

The first lawsuit regarding the overrepresentation of minority children in special education (i.e., EMR classes) involved Chicano children. In *Diana v. Board of Education* (1970), nine Chicano children, ages 8 to 13 years and attending schools in Monterey County, California, were plaintiffs (see Henderson and Valencia, 1985, for details). The pupils — all from Spanish-speaking homes — claimed they were inappropriately assigned to EMR classes on the basis of IQ

scores. Based on widely used IQ tests, the children's IQs ranged from 30 to 72 with a mean score of 64; all tests were administered in English. Upon retest in Spanish, seven of the nine children performed higher than the cutoff point for EMR placement; the other two children's retest scores were only a few points below the cutoff. The plaintiffs contended their placements were inappropriate because the tests administered (a) were standardized on White, native-born children, (b) contained cultural bias, and (c) placed heavy emphasis on English verbal skills (Weintraub and Abeson, 1972; cited in Henderson and Valencia, 1985). *Diana* was settled by consent decree and the final order contained a number of nondiscriminatory provisions — some that later would be part of federal law.

Following the *Diana* case, several other significant and similar cases involving Chicano, Black, and American Indian children were filed in California and Arizona. Each case had a 'piggy-back' effect in which new elements of logic and strategy were added — eventually helping to shape guidelines for psychoeducational assessment and services for Chicano and other racial/ethnic minority children. As Henderson and Valencia (1985) underscore, these lawsuits brought forth by minority plaintiffs proved to be extremely instrumental in molding the future of nondiscriminatory assessment. Out of this judicial crossfire evolved the eventual banning of group-based and some individually administered intelligence tests. The implementation of significant legislature reform (e.g., Public Law 94-142, the Education for All Handicapped Children Act of 1975 [*Federal Register*, 1977]) was also influenced by these lawsuits. In that test bias was a central issue in the preceding minority cases, it is not surprising that of the nine major mandates seen in PL 94-142, three are especially germane to minority children. That is, nonbiased assessment is required (meaning that evaluation and testing materials must not be racially or culturally discriminatory), tests and other evaluation measures must be validated for specific use, and if possible, psychoeducational assessment must be in the child's native language (Henderson and Valencia, 1985). In sum, the quality of the test instrument and how it is administered were key concerns in legislative reform. We now turn to a closer examination of the topics of test bias research and nondiscriminatory assessment during the post PL 94-142 years.

Test Bias Research and Chicanos

With the banning of group-administered intelligence tests and their pernicious sorting consequences, the focus of actual and potential test abuse with respect to Chicanos shifted to individually administered intelligence tests that were used in part for possible special education placement (e.g., EMR; learning disabilities; gifted and talented). The major question researchers asked was: Are the widely used individually administered, standardized intelligence tests biased against Chicano and other racial/ethnic minority children? For example, does the Wechsler Intelligence Test for Children — Revised (WISC–R; Wechsler, 1974) have differential predictive validity of academic achievement for White and Chicano students? Does the Kaufman Assessment Battery for Children (Kaufman and Kaufman, 1983) have differential construct validity for Whites and Chicanos?

Although research on text bias existed prior to the implementation of PL 94-142 (see Jensen, 1980 for a review of pertinent studies), it was not until the late 1970s and into the 1980s that bias research became a noticeable area of concern. Before discussing some of this research that is germane to Chicano

students, it is important to clarify a few terms and to provide a backdrop. Jensen (1980) notes the importance of making distinctions between the concepts of 'cultural loading', 'culture biased', and 'test unfairness'. Cultural loading, according to Jensen, basically refers to test items that '. . . consist of artifacts peculiar to a particular period, locality, or culture . . .' (p. 133) or are items that make use of school knowledge or skills (e.g., reading). Given this definition, all tests are culturally loaded to a certain degree. Cultural bias, which is bias involving racial/ethnic group membership, is concerned with psychometric bias.[11] As Jensen comments, this notion of bias is strictly statistical — that is it refers to the systematic errors (e.g., in the predictive validity) of test scores of individuals that are linked to group membership. As such, the assessment of cultural bias is purely objective, statistical, empirical, and quantifiable. On the other hand, the term test unfairness (and its reciprocal, test fairness) are subjective value judgments involving the use of test results (e.g., selection procedures).[12]

Since the advent of intelligence testing of minority children over seventy-five years ago, the issue of test bias and unfairness has been raised and addressed in a number of ways. During the 1970s, the test bias question concerning minority children was one of the most heated issues discussed in the educational assessment literature. Some scholars during that period contended without equivocation that conventional intelligence tests were biased against minorities (e.g., Alley and Foster, 1978; Williams, 1971). Debates on the definition of test bias and fairness became commonplace (e.g., Cleary, 1968; Green, 1975; Thorndike, 1971). Although the concept of test bias was still being debated in the 1980s, the current focus of mental testing research is primarily psychometric investigations of possible test bias in intelligence tests and other tests of mental ability. A number of scholars in the 1980s came to the conclusion that currently used intelligence tests generally are *not* biased against minority children. A good example of this view is the work of Jensen (1980), who states with strong conviction in his preface to *Bias in Mental Testing*:

> Many widely used standardized tests of mental ability consistently show sizable differences in the average scores obtained by various native-born racial and social subpopulations in the United States. Anyone who would claim that all such tests are therefore culturally biased will henceforth have this book to contend with.
>
> My exhaustive review of the empirical research bearing on this issue leads me to the conclusion that the currently most widely used standardized tests of mental ability — IQ, scholastic aptitude, and achievement tests — are, by and large, *not* biased against any of the native-born English-speaking minority groups on which the amount of research evidence is sufficient for any objective determination of bias, if the tests were in fact biased. For most nonverbal standardized tests, this generalization is not limited to English-speaking minorities. (p. ix)

In addition to Jensen, other researchers (e.g., Dean, 1980; Miele, 1979; Reynolds, 1982, 1983; Reschly, 1979; Sandoval, 1979) have provided evidence that certain intelligence tests (e.g., WISC–R) are not biased against minority children. This recent activity of test bias research and debate has generally been healthy for the testing movement, because issues of delineation, detection, and

BEST COPY AVAILABLE

minimization of bias in testing have been opened to wider, and even conflicting, perspectives (e.g., Berk, 1982; Bigelow, 1982; Gould, 1980; Jensen, 1980; Lambert, 1981; Reynolds, 1982; Reynolds and Brown, 1984; Reschly, 1979; Valencia and Rankin, 1985). Let us now move to a brief overview of test bias research pertinent to Chicano students, but first, this point of clarification.

Test bias research across racial/ethnic groups is different from single population validity investigations. Although validity coefficients are useful in the assessment of test bias, they are limited in scope. As Jensen (1980) notes, the concepts of validity and bias are separate questions. Whereas validity can apply to a single population (e.g., Valencia, 1984, 1985a, 1985b), the study of bias always involves a comparison of two or more populations — typically called 'major' and 'minor' groups. With respect to racial/ethnic minority students and test bias research, a study would require a major (e.g., White) and a minor (e.g., Chicano) group (see, for example, Valencia and Rankin, 1986, 1988). There can be, of course, variations of the major-minor group design in investigating cultural bias in tests. For example, in Valencia and Rankin (1985), the major and minor groups were English-speaking Chicano and Spanish-speaking Chicano children, respectively. In summary, the comparison of two or more racial/ethnic groups is the preferred strategy in investigating cultural bias in mental and other tests. Yet, although single population validity studies cannot examine cultural bias directly, they are still valuable in providing insights if validity coefficients are of sufficient magnitudes to conclude that an instrument has clinical utility for a particular minority group (e.g., see Valencia, 1988).[13]

In that the WISC–R is the most frequently individually administered intelligence test for school-age children, it is not surprising that this measure has also been one of the most heavily researched for racial/ethnic (i.e., 'cultural') test bias (Reynolds, 1983). Regarding WISC–R test bias investigations in which Chicano children have been compared to White children, there has been considerable research.[14] Because these studies have been reviewed elsewhere (e.g., Reynolds, 1982, 1983, and to some degree Jensen, 1980), we will only touch on the highpoints.

Test bias research on the WISC–R using Chicano and White samples has been conducted in the areas of construct validity, predictive validity, content validity, and reliability. With respect to construct validity, the basic approach compares the similarity of the factor structure of the WISC–R across Chicanos and Whites. If one finds that the test has factorial similarity for both groups, than it can be concluded, to some degree, that the WISC–R is not biased in construct validity. Investigations by Dean (1980), Gutkin and Reynolds (1980), Oakland and Feigenbaum (1979), and Reschly (1978) have supported the consistent similarity of WISC–R factor analyses across Chicanos and Whites. In short, based on these specific investigations (and of course, the populations sampled), one can conclude that the WISC–R measures the same constructs with approximately equal accuracy for Chicanos and Whites (Reynolds, 1982).

Chicano and White comparisons for potential bias in predictive validity have also been undertaken. In such studies, it is typical to test for homogeneity of regression across Chicanos and Whites (nonbiased prediction).[15] If statistical differences in the two groups' slopes, or intercepts, or standard error of estimates are found, then this would suggest bias in prediction if a common regression line

(Chicano and White combined) is used. Studies by Reschly and Reschly (1979), Reschly and Sabers (1979) and Reynolds and Gutkin (1980) have shown that the WISC–R does not have differential predictive validity (i.e., is not a biased predictor) across Chicanos and Whites.

On the subject of potential bias in content (item) validity on the WISC–R, Sandoval (1979) found the items x groups interaction to account for a very small percentage of the variance in WISC–R performance across Chicano and White children. Based on this investigation, it can be concluded that for the populations studied, WISC–R items were relatively not more difficult for Chicanos than for Whites. Finally, with respect to reliability estimates of the WISC–R across Chicanos and Whites, there is some evidence that the internal reliability of the WISC–R demonstrates an acceptably high degree of consistency for both groups (Oakland and Feigenbaum, 1979; Sandoval, 1979).

The preceding brief overview of WISC–R test bias studies indicates that this popular and widely used individually-administered intelligence test is, by and large, not psychometrically biased against *English-speaking, native-born*, Chicano children.[16] Does this mean, however, that *other* individually-administered intelligence tests are likewise nonbias when used with the Chicano school-age population? Although there is scattered evidence to indicate that some such instruments are free of bias with respect to Chicanos (see, for example, reviews by Jensen, 1980), the best answer to the above question is: given the paucity of test bias studies (i.e., investigations involving instruments other than the WISC–R), it is not possible to draw conclusions one way or the other. That is, in light of the very small number of non-WISC–R bias studies involving Chicanos and Whites, one simply does not know whether the instrument in question is bias or nonbias. Certainly, this pushes the point further along that research on test bias with White and Chicano populations is sorely needed.

The necessity for vigorously undertaking test bias research regarding Chicano children has been recently underscored in a series of studies that provide evidence on the complex nature of test bias findings and interpretation (Valencia and Rankin, 1986, 1988, 1990). The subjects in the Valencia and Rankin investigations were White and Chicano fifth- and sixth-grade boys and girls, and the instrument under examination for potential bias was the Kaufman Assessment Battery for Children (K–ABC). The K–ABC contains an intelligence scale (Mental Processing Scale) and a separate Achievement Scale. Four investigations of possible test bias (against the minor group — Chicanos) were undertaken by Valencia and Rankin. Using a variety of test bias statistical analyses to study potential bias in the K–ABC along lines of three types of validity (construct; content, i.e., item; predictive) as well as reliability, some mixed results were found. Bias *was not* found in construct validity and reliability (Valencia and Rankin, 1986), but bias *was* identified in content validity (Valencia and Rankin, 1990) and predictive validity (Valencia and Rankin, 1988).[17] In the predictive validity study, Valencia and Rankin (1988) offer this conclusion regarding the complicated K–ABC bias findings:

> . . . the K–ABC appears to be flawed or biased when used with Mexican American students, because the test does not have the same predictive efficiency with the majority and minority students. Finally it is worth

> noting, as does Jensen (1980), that a test could have the same degree of construct validity in both the majority and minority groups (i.e., as seen with the K–ABC in Valencia and Rankin, 1986), even when the predictor variable has comparable reliabilities in both groups (i.e., as seen in Valencia and Rankin, 1986 ...), and yet be a biased predictor of achievement. (p. 263)

It is important to keep in mind that the above 'mixed bag' of test bias research on Chicanos is only a small sliver of the potential research that could be undertaken. It is important to keep in mind that evidence presented in the Valencia and Rankin studies is for *one* instrument (K–ABC), *one* age group (11-year-olds), *one* location in *one* state (central California city), and so forth.[18] It is not difficult to imagine that given the number of intelligence tests available and the variation in possible test bias research focuses and designs, there are indeed a large number of research investigations that are in the realm of possibility.

Unfortunately, at a time when more research is needed on test bias with Chicanos and other racial/ethnic minorities, there appears to be a gradual decline in interest — and thus investigatory activity — by researchers. Peaking in the late 1970s and early 1980s, test bias research (especially on individually-administered intelligence measures used in the assessment of school-age children) began to decrease in the mid 1980s and even into the 1990s. This nosedive in research activity and publication could be related, in part, to the work of some scholars who may have helped close the door to test bias research by prematurely drawing broad conclusions that most mental measures are relatively free of cultural bias (e.g., see Jensen, 1980). A second contributing factor to the decline in test bias research is likely related to the very nature of test bias. As Reschly (1979) comments, test bias is an issue filled with emotion — and for a long time at that. To some degree, the contemporary period from the late 1970s to the present can be viewed as a moody reflection of the public and scholarly waxing and waning toward the test bias controversy seen over the decades. Third, perhaps the decrease in test bias research is connected to a more general climate of apathy toward minority children in the US. For example, in speaking to the need for further theoretical understanding with respect to the widespread issue of poor schooling performance of minority children, Boykin (1986) observes: 'The question is particularly crucial today, at a time of declining political interest in minority affairs. Minority children no longer enjoy national attention, but their educational problems persist' (p. 57).

Whatever the probable reasons might be for the diminished attention to test bias research, it is certainly not because the measurement community lacks the technology. From the advent of mental testing up to the 1960s, there was considerable confusion in test bias research and literature. According to Jensen (1980), much of this disorder was due to inconsistencies in terminology and a lack of clarity in conceptualizing and differentiating bias (a statistical notion) and unfairness (an abuse of test results). In the last ten or so years, however, a number of publications have appeared discussing statistical and methodological approaches for measuring test bias. This body of research — in both theoretical treatises and actual empirical investigations — has greatly enhanced the state-of-the-art of test bias conceptualization, methodological detection, and interpretation (see, for example, Berk, 1982). In short, given what current measurement

technology can offer, it seems to us that there is no excuse for Chicano and other minority children to be tested on mental measures that have not been scrutinized for potential cultural bias. And, of course, it would be unpardonable to administer tests to children in which cultural bias has been identified (e.g., see Valencia and Rankin, 1988).

The Responsibility of Test Publishers

There is no doubt that the 'stuff' of the measurement community — that being psychometric theory, knowledge, and application — is continually needed to push along the realization of nondiscriminatory testing and assessment. But, what about the role of test publishers — those who develop and market tests. Should they also have a responsibility to ensure that their products are free of cultural bias and to help prevent the abusive practice (albeit frequency unknown) of examiners who administer tests with poor or unknown psychometric integrity? To even get a sense of this issue, we need to go beyond the topic of individually-administered intelligence tests and into the broad field of test publishing.

Mitchell's (1984) paper on 'Testing and the Oscar Buros lament: From knowledge to implementation to use' provides an incisive look into the many problems of published tests. As Mitchell notes, Buros (the founder of *The Mental Measurements Yearbook*, the world's richest source on the quality of published tests) years back had some harsh words on tests. In the *Eighth Mental Measurements Yearbook* (*MMY*; Buros, 1978), Buros charged, that by and far, the publishers of tests continue to sell tests that do not meet the minimal standards of the *MMY* and test reviewers. According to Buros, 'At least half of the tests currently on the market should never have been published' (see, Mitchell, 1984, p. 113).

Mitchell (1984) cites a small descriptive study conducted by the Buros Institute of Mental Measurements staff that affirms Buros' lament about published tests. The test reviews in *The Eighth Mental Measurements Yearbook* were examined in order to see how well test publishers attended to providing critical test data (i.e., evidence of reliability, validity, and norms). The results of this investigation were discouraging to Mitchell. Key findings were:

1 As a whole, about 41 per cent of the tests listed in the *MMY* '... were lacking reliability and/or validity data in some important respect. Tests in the areas of reading, vocations, and speech and hearing were the worst offenders' (pp. 114–15).
2 Regarding norms, 'All told, 28 per cent of the tests listed ... were inadequately normed in some important respect' (p. 115).

A few other specific points that Mitchell (1984) cites about test publishing further support Buros' disappointment with the quality of tests. To wit,

3 There has been a proliferation of tests. Based on an analysis of *Tests in Print II* (Buros, 1974), there are 496 test publishers listed. Although less than 2 per cent of the publishers publish 26 per cent of all tests, Mitchell notes that the majority of publishers (58 per cent) have just a single test listed, about 75 per cent have three or less tests, and 85 per cent have five or fewer. Mitchell concludes:

> . . . there is much of the cottage industry to the test publishing business, and there are many test publishers distributing their own tests or very small test publishers with single or extremely limited test offerings or book or instructional materials who have acquired a few tests and publish them in a manner almost incidental to their major interest and thrust. (p. 113)[19]

The impact of such an uneven industry is quite revealing. First, according to Mitchell, there is the fact that more and more tests are being published, but of poorer and poorer quality. Second, the developmental and marketing costs — and of course, the huge profits of sales from poor and marginal tests — all indicate that the measurement community is losing ground in trying to maintain some semblance of test quality. Third, as more poor tests appear on the market, it becomes more difficult for test users and the public to become discriminating consumers.

4 Claims about validity evidence are often overstated. Mitchell (1984) notes that it is not uncommon for test publishers to shunt aside modest to weak validity evidence and to create illusions than a great deal more benefits from tests can be offered. As such, it is irresponsible for test publishers to promote test utility in the absence of strong validity evidence.

With the broader issue of 'Buros' lament' in mind, it becomes exceedingly clear that test publishers indeed have a major responsibility and challenge in developing nonbiased tests (intelligence and otherwise) when such instruments are to be used in the assessment of Chicano and other culturally and/or linguistically diverse populations. Reynolds (1982) fittingly describes this mandate as such:

> Test developers are . . . going to have to become more sensitive to the issues of cultural bias to the point of demonstrating on publication whether their tests have differential content, construct, or predictive validity across race or sex *prior* [italics added] to publication . . . With the exception of some recent achievement tests, *this has not been common practice* [italics added], yet it is at this stage where tests can be altered through a variety of item-analysis procedures to eliminate any apparent racial or sexual bias. (p. 208)

Summarizing matters thus far, we have provided some historical and contemporary insights to the uses and abuses of intelligence testing with respect to Chicano students. Historically, there is some rather convincing evidence that group-based intelligence testing was used, in part with other educational practices, to help shape limited educational opportunities for Chicanos, both in the educational mainstream and special education. With the demise of group-administered intelligence tests, the contemporary spotlight is now on individually-administered intelligence instruments with respect to potential test bias and the abusive practice of using psychometrically poor tests. As we have discussed,

despite a substantial technology the measurement community has failed in recent years to muster enough continued interest and energy to pursue the needed research into the question of test bias. By no means is the issue of test bias with respect to Chicanos and other racial/ethnic minority groups a closed issue. The resolution of the test bias question lies in the 1990s, and perhaps beyond.

The researcher, however, does not stand alone in his or her responsibility to bring further light to the question of potential test bias. We have seen that the test publishing community also needs to share in providing bias-free psychoeducational instruments. The Buros lament becomes even graver when one includes the issue of potential test bias. Certainly, for test publishers not to meet minimal standards of adequate reliability, validity, and norming in their educational tests is unjustifiable. For such publishers to also market their poor or marginal quality tests without having undertaken investigations to ensure nonbias across racial/ethnic groups is unconscionable. Indeed, in view of the major abuses at the test publishing level, 'a call to action' for the improvement of test development is in order (see Mitchell, 1984, for several recommendations).

The Responsibility of School Psychologists

In addition to the research/measurement and test publishing communities, there is also a third, significant sector that has responsibility in insuring nonbiased testing and assessment — the practitioners (school psychologists). With respect to intelligence testing, an important traditional task of the school psychologist has been to test children who may be suspected of functioning subnormally in intellectual behavior. Nowadays, this role of the school psychologist presents some serious problems. Henderson and Valencia (1985) capture the issues this way:

> School psychologists were able to go about that task confident that the instruments they used were reliable and valid for their use. The intelligence test, sometimes thought of by educators as a sort of appendage to the school psychologist, was widely regarded as the single most impressive achievement of psychological science. Today, school psychologists are less confident of their assessment tools, ensnarled in an ethical and professional dilemma. (p. 340)

The ethical and professional fix in which school psychologists find themselves is this: on one hand, school psychologists are required by law (e.g., PL 94–142) to search actively for and identify children who may need special educational services or programs. On the other hand, the same legislative mandate requires nondiscriminatory assessment. As these dual (but obviously connected) responsibilities have intensified, school psychologists find themselves more and more troubled (Henderson and Valencia, 1985). Traditional assessment tools, especially the individually-administered intelligence test, are increasingly being called into question for use with children whose racial/ethnic or social class backgrounds fall outside the modal configuration (i.e., English-speaking, middle-class White mainstream).

As such, school psychologists are placed in a difficult position. They need to scrutinize their test arsenals by asking two vital questions. First, do the tests meet

the minimal standards for acceptable levels of reliability and validity, as well as for appropriate norming? Second, and an extension of the first concern, is there evidence that the tests in question are free of cultural bias along lines of reliability and the various types of validity? If the answers to these questions are 'no', and if a particular test is administered nevertheless, in our opinion this would constitute an abusive practice of testing.

Inappropriate and deleterious testing practices are not only confined to administering tests that lack good psychometric quality (in general, and with respect to nonbias). First, abuses can also occur by school psychologists when administering intelligence tests, or other various types of educational tests, to students who are not proficient enough in English to handle the verbal demands (see Valencia, 1990, for a discussion of how inattention to this issue has created and continues to create, assessment problems for Chicano students; as well, he presents an overview of how methodologically confounded intelligence testing research with Chicanos was shaped because of the language issue). Second, there is the problem of not attending to multiple data sources of assessment. Salvia and Ysseldyke (1988) underscore that '. . . testing and assessment are not synonymous' (p. 5). Unfortunately, the belief that testing and assessment are identical has led some school psychologists astray and has caused inappropriate diagnosis and intervention for some students, especially minorities. There is, however, some optimism as the field of school psychology has gradually become more sensitive to the need for expanded ways and models to be used in identifying and providing services for children with learning difficulties (e.g., see Oakland and Goldwater, 1979, for a discussion of assessment and intervention models for mildly retarded children; see Henderson and Valencia, 1985, for a general discussion on expanded assessment and intervention strategies for minority children).

A third area of potential abuse in the testing of Chicanos and other minority students has to do with the failure of school psychologists to *go beyond* bias-free testing. There is no doubt that in the work of school psychology, nonbiased testing is crucial. But, as Henderson and Valencia (1985) contend, 'nonbiased testing is useless unless it results in nondiscriminatory education' (p. 342). It is important for the school psychologist (in a consultative model) to work with teacher and parent in connecting testing and assessment to instructional services and programming.

Along these lines, Henderson and Valencia (1985) argue that despite the limitations of current tests and practices, nondiscriminatory psychological assessment and services are attainable if school psychologists adhere to certain principles. Some examples are:

1. Be knowledgeable about the cultural backgrounds of children and the demand characteristics of the home and school environments.
2. Function as problem solvers. That is, be open to multiple sources of assessment data. Be sensitive to possible cultural influences on the performance in question. Use the gathered data at hand as the basis for testing hypotheses in the context of intervention.
3. Employ a 'consultation' model rather than the traditional 'refer-test-report' model. For example, under the former model, the school psychologist looks for the underlying causes of learning and remediation in the

BEST COPY AVAILABLE

classroom, rather than in superficial analysis and intervention through potentially meaningless contingency management.

4 In the absence of well-trained psychologists (especially the absence of minority school psychologists), assessment and service delivery could profit from having a cultural informant or ombudsperson.

Beyond Nonbiased Assessment

In closing this section on intelligence testing, we wish to expand the preceding discussion on the relation between nondiscriminatory testing (and assessment) and nondiscriminatory schooling. We do this by placing the discussion in the context of a major focus of our discussion thus far, that is, test bias. Conceptualizing test bias as a psychometric notion has proven to be a valuable contribution to the measurement and interpretation of bias. This contribution is largely so because the psychometric approach defuses, to some degree, the emotional debate often associated with the question of bias. Furthermore, this approach can and does lead to empirically defined and testable definitions of bias (in the case of Chicano students, see for example, Valencia and Rankin, 1985, 1986, 1988, 1990). In recent years, however, reservations have been voiced about the strict statistical approach to understanding bias. The major criticism of statistical, or psychometric, test bias is that it is rather exclusive in its conceptualization. One sub-criticism of this exclusivity area is that there is a muddling of the terms bias and unfairness — a topic we discuss next.

As discussed previously, a number of scholars (e.g., Jensen, 1980) make a sharp distinction between 'bias' and 'unfairness'. While bias and nonbias are empirical and statistical matters, unfairness and fairness are moral and legal issues dealing with how test scores are used in a selection situation. According to Shepard (1982), the distinction between bias and unfairness, however, is problematic. First, the intended difference between the two terms is not unequivocally conveyed in everyday communication. Shepard notes: 'To be biased is to be unfair, unjust, prejudiced. Calling your test "biased" conveys nearly the same message as a placard calling an employer "unfair"' (p. 10).

Second, the distinction between the notions of bias and unfairness presents some awkwardness in the psychometric sense. Shepard (1982) notes that although authors generally conceptualize bias as a form of invalidity, bias '. . . is now being taken as an inherent feature of a test, while its opposite, validity, has always been considered to be a property of test use, not of the test itself' (p. 10). An example of this disorder is that the difference in the two terms is confused by the bias-in-selection literature (see Peterson and Novick, 1976; cited in Shepard, 1982), which definitely pertains to test use but suggests different models of predictive validity as bias indicators. In this context, bias is thought of as a specific type of invalidity that is *outside* of the test instrument rather than being an inherent feature of the test itself. Notwithstanding the confusion between bias and unfairness, Shepard argues that it is worthy to maintain the distinction intended by some authors, because the difference between bias in tests and unfair test use is critical to an understanding of bias detection. Shepard (1982) does, however, raise this helpful suggestion:

> It is possible to be faithful to the rule that validity must always pertain to the particular inferences made from a test, yet still admit different

> degrees of externality through which bias may be more or less closely associated with the *use* of a test, rather than its internal characteristics. There is a validity continuum, anchored at one end by unbiased tests that measure what they were designed to measure and do so equally well for all groups. Further along the continuum are tests that provide equal predictive validity in particular contexts. At the farthest end of the continuum are tests for which validity is established by resolving issues of justice and values, as well as scientific arguments over what statistical model of fairness to apply and what the criterion will be. (p. 11)

Given the obviously different points of view, what definition or conceptualization of test bias should be adopted? Henderson and Valencia (1985) ask: 'Is it best defined as the poor match between test content and cultural experiences of minority and poor children, the influence of situational factors on performance, or technical validity?' (p. 350). Regarding the latter (i.e., statistical conceptualization), there is some value in it because, as Shepard (1982) notes, it helps us to study and understand how to detect bias. Furthermore, as we have discussed earlier, such an approach has greatly assisted in providing some evidence that the more well-developed tests (e.g., WISC–R) are relatively nonbias. The existing corpus of statistical research on test bias is significant in advancing our knowledge base and certainly needs to be encouraged. Yet, as we have underscored, the test bias question with respect to Chicano students and other minority groups is far from being resolved.

Notwithstanding the value and contributions of the statistical paradigm in identifying and measuring cultural bias on individually-administered tests (as well as other psychoeducational instruments), we think it is misleading to adhere to such an exclusive conception. Bersoff (1984), Henderson and Valencia (1985), Messick (1989), Reschly (1979), Shepard (1982) and others have come to the heart of the matter with the observation that *tests do have social consequences*. Since the seminal period of testing, tests have been used to open doors for some, and close doors for others. Abusive school testing practices *vis-à-vis* Chicanos, other minorities, and the poor often constitute institutional racism in which self-perpetrating, unquestioned testing practices within school systems diminish learning opportunities for children (Henderson and Valencia, 1985).

In sum, the subject of test bias is complex and controversial. In our opinion, as well as others, it is scientifically and ethically inappropriate not to tie testing/assessment with schooling and its consequences. The statistical notion of bias is useful, but alone it has little meaning. On this, Bersoff (1984) voices: '... reliance on psychometric models for test bias without consideration of the social and ethical consequences of test use ignores the concerns of significant segments of society' (p. 105). Or, as Reschly (1979) points out: '... to defend tests on the basis of evidence of common regression systems or to attempt to separate the issues of technical adequacy from those of social consequences is insufficient' (p. 235). In the final analysis, the administration of intelligence *and* other educational tests to Chicano students should be considered in the context of the broad institutional processes that help shape school problems for them. The wider implications of testing/assessment and the schooling context for Chicanos need to be considered in any effort to envision nondiscriminatory school services.

Reschly (1979) grasps the fundamental nature of this matter in his assertion that

> The ultimate criteria that should guide our evaluations of test bias are the implications and outcomes of test use for individuals. Succinctly stated, test use is fair if the results are more effective interventions leading to improved competencies and expanded opportunities for individuals. Test use is unfair if opportunities are diminished or if individuals are exposed to ineffective interventions as a result of tests. (p. 235)

Competency Testing

As we previously discussed, the issue of 'accountability' in education is a widespread concern in our society. The current 'competency testing' movement is part of the broader public demands for accountability in the nation's schools. In a nutshell, the notion of competency testing carries with it a gatekeeping function in which examinees are tested to see who will be promoted, graduated, admitted, or certified. In this section we begin with a description of the types of competency testing. Following this is a discussion of minimum competency testing and teacher competency testing and their negative impact on Chicanos.

Types of Competency Testing

In our conceptualization of the nature of competency testing, we see three broad forms. First, there is 'minimum competency testing' (MCT). As Jaeger (1987) notes, MCT began in a big way in the early 1970s. Oregon's State Board of Education mandated the school systems in 1972 to develop and implement a statewide program to measure student 'competence' (also see Herron, 1980). Soon after, MCT spread nationwide. Baratz (1980) comments that by 1977, eighteen states jumped into the MCT movement. By 1987, forty states had climbed the bandwagon (Jaeger, 1987). It appears that a primary use of MCT is to award or deny students a high school diploma. After nearly a decade of MCT, in 1980 about 50 per cent of the states used the passage of a test of minimum competence as a prerequisite for the earning of a diploma (Lerner, 1981; cited in Bersoff, 1984). As we enter the 1990s, many states use MCT in such a way. Based on what little data are available, Chicanos and other racial/ethnic minority student groups have higher failure rates than their White peers on these measures, and thus are denied high school diplomas at higher proportions. Later, we will return to the controversial MCT and the resultant abusive practices.

A second major type of competency testing is what we call 'school-based competency testing'. This form of testing typically involves a required statewide system of student evaluation of minimum skills in basic schooling areas (e.g., reading). These mandates are top-down in which state legislatures commonly require school districts throughout the state to administer broad-based achievement tests (e.g., mathematics, reading, writing) to elementary and secondary students (usually at selected grade levels). It is typical for such testing programs

to be part of an omnibus school 'reform' package passed by the legislature. Examples of school-based competency testing are the California Assessment Program (CAP) and the Texas Educational Assessment of Minimum Skills test (TEAMS). Typically, a state educational agency provides, after testing, all school districts with aggregated test results (e.g., the unit of analysis is the individual school).[20] In some states, the actual testing, analysis, and reporting of results has spawned bureaucracies. For example, a recent report of TEAMS results in Austin, Texas, contained great detail of test score comparisons for a section of the local district (Christner and Moede, 1988–89). TEAMS scores were compared between schools, grade levels, and racial/ethnic groups.[21]

In sum, it appears that the results of school-based competency testing are filling an accountability function of a sort. By far, compared to the other types of competency testing, school-based testing results are made the most public. It is quite common for local newspapers to report, in some detail, the outcomes of local testing. Depending on the scores of these 'public report cards', local superintendents — or in some cases the highest ranking state public educational official — may flaunt the test results, contending that 'real progress' is being made and tests score 'are up' from the previous year (Phillips, 1989; Watson and Kramer, 1989). Sometimes, school officials will even discuss problems, underscoring the percentage of students (and often naming the schools) who have 'failed' the competency testing (Graves and Breaux, 1989). We do not find fault with the public's right to know the results of school-based competency testing. We do, however, find disturbing, misleading, and abusive the manner in which test reporting is done in some instances. For example, in some cases school officials in bi- or multiracial communities will report test results in such a way that minority parents are misled to believe that their children are achieving at satisfactory levels.[22]

The third form of competency testing can be placed under the rubric of 'teacher competency testing'. As another child of the parental accountability movement, the teacher competency testing movement began in 1978 and has now swept the country (Valencia and Aburto, in press, a). The term 'teacher tests' is an umbrella for three forms of paper-and-pencil teacher competency tests. An 'admissions' test is a basic skills test required as an entry criterion to a teacher education program. A 'certification' test is also a basic skills test and/or a professional knowledge test and/or a subject matter test required as a condition for earning an initial teaching credential granted by the state. A 'recertification' test is a basic skills test required of incumbent teachers. Based on the most recent data there are twenty-four states that require some type of teacher competency test for admission to a teacher education program; thirty-six states require such testing only upon graduation as part of state certification, and eighteen states mandate both entrance (admissions) and exit testing (certification) (Eissenburg and Rudner, 1988). There are three states that require teacher competency testing for teachers currently practicing (recertification) (Shepard and Kreitzer, 1987).

Approximately a decade ago, concern was raised at public, political, and school levels about the preparedness and effectiveness of beginning teachers. Because of the continuing criticisms of America's teachers and schools (e.g., *A Nation at Risk* by the National Commission on Excellence in Education), the public is demanding some assurance from its state agencies that teachers who become licensed are actually competent — hence the introduction of competency

tests. The central idea behind such testing is that before people be allowed to teach, they must demonstrate 'basic skills' (e.g., mathematical ability, reading, writing) that are believed to be necessary to carry on day-to-day instructional activities. Although the motive underlying teacher competency testing is clear (i.e., the need to upgrade teacher quality), the nature of teacher testing is fraught with conceptual, measurement, and social problems. In the case of prospective Chicano teachers, they (and other racial/ethnic minority groups, particularly Blacks) have been forced to carry a very disproportionate burden of teacher reform efforts. That is, Chicanos and other minority examinees, compared to their White peers, have failed teacher tests at very high rates — to such a degree that the minority teacher shortage is at a crisis situation (Valencia and Aburto, *in* press, a). As will be elaborated later, the sharp decline of Chicano and other minority teacher comes at a time in which the minority school-age population is growing at dramatic rates.

In short, competency testing with respect to Chicanos is filled with controversy. Specifically, these issues include: (a) conceptual confusion about the distinction between 'competence' and 'incompetence' (e.g., Jensen, 1980); (b) arbitrariness and scientific indefensibility of standard setting (i.e., arriving at a cut score to determine who passes, who fails; (see Valencia and Aburto, in press, b); (c) 'high-stakes' nature of competency testing in that one's future rides on a single score (e.g., Madaus, 1986); (d) dire social and educational consequences for Chicano examinees and Chicano school children (e.g., Valencia and Aburto, in press, a). Given the paucity of research in the area of school-based competency testing and because of space limitations, the above issues will only be examined in the areas of MCT and teacher competency testing.

Minimum Competency Testing

As part of major educational reform efforts aimed at improving the quality of our schools, many lay boards of education and state legislatures have responded over the last sixteen years by implementing MCT programs in their states. MCT is primarily based on a belief that testing of essential skills and competencies (e.g., math, reading, and writing) will help raise academic standards, increase educational achievement, and restore public confidence in education. The passage of a minimum competency test for high school graduation and/or grade-to-grade promotion is currently required in at least forty states (Haney and Madaus, 1978; Paulson and Ball, 1984). A 1985 *Education Week* survey indicated twenty states require high school students to show mastery on a state-mandated exit test as a prerequisite to receipt of a regular high school diploma (Airasian, 1988). In this section, we will discuss: (a) criticisms of MCT, (b) adverse impact on Chicano and other minority students, and (c) the effect of MCT on the operation and structure of schooling.

Criticisms of MCT

Boasting strong public and political support, MCT proponents contend that students will benefit by mastering the basic skills, raising their self-confidence, and enhancing their career opportunities. While the MCT movement addresses a variety of social and political purposes, its most widely recognized goal is the improvement

of students' basic skills. Ideally, the test's main function is to identify deficiencies that may be treated through remediation. A critical feature in some programs is the use of test results as a screen for high school graduation (Serow, 1984). In general, students are required to pass a test demonstrating 'minimum competency' in basic academic skills and their practical application to 'real-life' demands before receiving their diplomas (Jensen, 1980). Supporters argue that racial/ethnic minority students will particularly benefit from MCT because it will reveal inequities in their education so they may be rectified (Paulson and Ball, 1984; Serow, 1984). In that most students are believed capable of attaining the requisite level of competency before graduation day, the diploma sanction is not considered by some as an act of discrimination against Chicanos, other minority students, or the poor (Serow, 1984).

Yet while competency tests appear simple, straightforward, and are widely used, their overall quality and intentions have been criticized from the start. Numerous writers view the almost exclusive reliance on MCT for awarding a high school diploma, for determining grade-to-grade promotion, or for assigning students automatically to remedial classes as classic examples of improper uses of tests. If in fact the tests were established in recognition of problems in the educational system, then to withhold diplomas, for example, simply punishes the victim (Linn *et al.*, 1982). Opponents also contend competency tests and standards serve more as short-sighted symbolic and political gestures than instrumental reforms (e.g., Airasian, 1988; Ellwein, Glass and Smith, 1988) and simply represent another area of potential discrimination against minorities and the poor, creating an additional obstacle to the attainment of social and economic equality in American life (Serow, 1984). Other critics sum up MCT as simply an effort to legislate educational success without concern for methods or modes of achievement (Jaeger and Tittle, 1980). Finally, there are some scholars who believe that because 'competence' is such a relative concept, it makes no psychometric sense to dichotomize students as being either 'competent' or 'incompetent' (Jensen, 1980).

Unfortunately, a brief survey of recent literature in the area shows that many of the early criticisms and fears raised over competency testing remain unresolved and few expectations or promises have been fulfilled. We now move to one major criticism of MCT — negative impact on Chicanos and other minorities.

Adverse Impact on Chicano and Other Minority Students

Of the many criticisms leveled against MCT, several relate to questions concerning its impact on minorities. From the start, opponents argued that MCT programs posed substantial risks for students from racial/ethnic minority backgrounds, who, through no fault of their own, experience school failure. Rather than forcing those students who fail competency tests to take their education more seriously, some critics believed the diploma sanction would instead lower students' motivation for attending schools, thus causing increased academic and disciplinary problems and higher dropout rates (Serow, 1984). Unless tied to an effective remediation program monitored by sensitive administrators, MCT could also be used to justify a new sort of segregation by placing failing students with the worst teachers or in less effective curriculum tracks (Paulson and Ball, 1984). Above all, competency testing further reinforces a stigma of failure for low-achieving students, and in the long run perpetuates racial and economic inequality (Serow, 1984). Though few studies have

examined the consequences of MCT it appears many negative premonitions have come to pass.

Failure rates on minimum competency tests for minorities, particularly Blacks, Chicanos, and other Latinos, are much higher than they are for White students. For example, early MCT trial run information from Florida showed that of 115,901 students taking the state's MCT in 1977, 36 per cent failed. Of those who failed, 78 per cent were Black, even though Blacks constituted only about 20 per cent of those taking the exam (Paulson and Ball, 1984). In 1978 it was reported that some 77 per cent of Blacks, 39 per cent of Latinos, and 24 per cent of Whites 'failed' the arithmetic test; furthermore, 26 per cent of Blacks, 7 per cent of Latinos, and 3 per cent of Whites 'failed' the reading test and writing portions (Jensen, 1980). Those students failing received a certificate of completion rather than a standard diploma. This is a significant impact when one considers that a certificate of completion is not considered a diploma for purposes of employment in the state of Florida or for purposes of admission to one of Florida's nine state universities. It was estimated that the denial of a diploma to Black students who failed the competency test resulted in a 20 per cent decline in Black enrollment in the state's universities and colleges (Paulson and Ball, 1984). Competency test performance data from the states of California, Florida, North Carolina and Virginia, also confirm the expectation that minority students experience greater difficulty in passing such tests than Whites (Serow, 1984). Supplementary data from the state of North Carolina revealed that students from lower-socioeconomic status (SES) backgrounds were about one-third less likely to pass the exam on their first attempt than students from higher SES backgrounds (Serow). As is well known, racial/ethnic minorities are usually concentrated in the lower SES categories.

Other studies have shown that even for students who stay in school, many fail to meet course and proficiency test requirements established at the district level. In 1981–82, minority high school seniors in California were three times less likely to complete the course requirements for graduation than were other students. Among those students that completed the district's course requirements for graduation, racial/ethnic group differences existed in passing the proficiency examinations required for graduation. As of December 1981, 17 per cent of the White students in this category did not pass one or more of the proficiency tests required for high school graduation, compared with 36 per cent of Black seniors and 25 per cent of Chicano and other Latino senior students (Brown and Haycock, 1985).

The effect of the MCT movement may be especially adverse for students who have previously suffered school failure because it places higher academic demands on those already at risk of dropping out (Archer and Dresden, 1987). It is difficult to distinguish between students who would have dropped out regardless of MCT requirements and those who became dropouts specifically because of their failure to pass competency tests. Data from Archer and Dresden's study, however, suggest that a significant number of Texas students already at a late junior or senior level will not receive high school diplomas as a result of failing the minimum competency test. This situation creates a new kind of dropout — students with a poor academic background who have the willingness to stay in school and graduate, but who will be denied a high school diploma because they do not meet the minimum standards.

Statistics from a 1985 administration of the Texas Educational Assessment of Minimum Skills (TEAMS) test taken their junior year indicates 12 per cent (22,485) of the examinees failed the mathematics section and 9 per cent (16,921) failed the English language arts section (Archer and Dresden, 1987). The demographic data suggest that failure to master the tests and attain diplomas will be disproportionately high for some groups. Those most effected by the exit level requirement will be students with limited-English proficiency being served in bilingual programs (48 per cent failing in language arts on their first try) and 'disadvantaged' students in Chapter I programs (39 per cent failing in mathematics). Blacks failed the mathematics portion at a rate of 28 per cent, and the failure rates for Chicano and White students were 18 per cent and 6 per cent, respectively. Language arts failure rates were 19 per cent for Blacks, 16 per cent for Hispanics, and 4 per cent for Whites. Additional students at risk of not receiving a diploma were almost 12,000 students who did not sit for this examination or any subsequent make-up administrations. Failure to master either subtest results in failure to receive a diploma. Unfortunately, this group of students will only join others from economically disadvantaged and racial/ethnic minority backgrounds who already traditionally have disproportionately high rates of truancy, dropping out, and school discipline problems (McDill, Natriello, and Pallas, 1985).

Very few states provide information on final diploma denial figures, especially with respect to the racial/ethnic or socioeconomic backgrounds of the pupils. The Texas data suggest, however, that MCT diploma sanctions are imposed disproportionately on Chicano and Black students (Archer and Dresden, 1987). Projections from California based on 1981 test results similarly show that Chicano and Black students are overrepresented among potential diploma denials (Brown and Haycock, 1985). Available data probably underestimate minority students' share of diploma denials because they fail to account for retest performances (Serow, 1984). North Carolina figures based on final results show Blacks made up about 25 per cent of the 1980 graduates but received more than 75 per cent of all diploma sanctions. In all, 4.4 per cent of all Black graduates were denied a diploma because of MCT failure, compared to 1.8 per cent of other minorities, and .5 per cent of White graduates (Serow, 1984).

Effect on the Operation and Structure of Schooling

At the onset, the hurried implementation of MCT could only allow for speculation about its effectiveness. Yet, fifteen years later, expediency is no longer a viable excuse for not knowing the effect of MCT on the operation and structure of our schools. Even a cursory search through the literature, however, shows that attention is still primarily focused on questions of competency definition, test development, standard setting, and program operations. Ellwein *et al.* (1988) note that over 60 per cent of the MCT research published between 1977 and 1987 is primarily rhetorical. Only 10 per cent of the entire literature could be classified as systematic, empirical research. Two perspectives examining the benefits of MCT reform are presented here. One discusses the effects of MCT on multiple school sites across the nation from an empirical perspective, and the other is a teacher's personal lament on the status of MCT in one school. Both indicate that the amount of attention given to minority issues or the remedial efforts aimed at offsetting the poor showing of all students who fail competency tests appears to

be minimal. We seem to have accepted these high-stakes examinations on what Airasian (1988) calls 'symbolic validation' without demanding that they justify their existence on any sort of empirical ground.

Aware of the need for an empirical view of MCT and standard-setting, Ellwein *et al.* (1988) studied five sites across the country. These researchers were guided by three major questions: For whom and for what purposes are test standards set? How and by whom are standards established? What consequences follow from the setting of standards? Results from the Ellwein *et al.*, investigation led to the formulation of five propositions that address the nature of competency testing reforms. Of special interest are the propositions noting (a) lopsided organizational activity, (b) nominal attention to minority issues witnessed at all five sites, and (c) the assertion that competency tests and standards function more as symbolic and political gestures than instrumental reforms.

Ellwein *et al.* (1988) note that organizational efforts are most visible, intense, and detailed during early phases of competency testing reforms — with similar efforts conspicuously absent in later stages. Intense efforts are concentrated in the development and administration of the instruments, but the amount of time invested contrasts sharply with the attention given to gauging and evaluating the effects of what the tests produced. In general, sites routinely figured only initial pass rates, with only a few collecting and reporting ultimate pass rates. Planned evaluations in general did not go beyond the tangible, technical outcomes of such rates.

The study also found that agency attention to minority issues is most prominent in efforts to build unbiased tests and most inconspicuous in efforts to assess adverse impact. In spite of differential pass rates at each site, issues of impact beyond the test themselves were left unexamined. Efforts for the most part centered on judgmental and technical reviews of items with no measure of the tests' impact. Only one site kept track of failure rates, number of repeat attempts, and the number and type of test-related decisions concerning retention or graduation. Given that some sites have large racial/ethnic minority student populations, failure to report such obviously important data is cause for concern (Ellwein *et al.*, 1988).

Of special importance is the conclusion by Ellwein *et al.* (1988) that '... competency tests and standards function as symbolic and political gestures, not as instrumental reforms' (p. 8). Two of five observed themes underpinning this conclusion are the loose coupling of test performance and subsequent decisions and the striking contrast between early and late phases of competency testing reforms. While information on the numbers of students who fail and do not graduate may exist within local institutions, such figures are not calculated or available at the state level and thus are not a matter of routine or public information. In addition, the attention given to test development, standard-setting, and general implementation contrasts sharply with the lack of attention to questions of impact, utility, and the value of competency tests and standards. In short, there seems little to say about the instrumental value of MCT and standards because it simply has not been examined in any of these sites. A lack of attention to these larger issues only further conceals any instrumental benefits or dangers these reforms may bring (Ellwein *et al.*, 1988).

A second reality check on what MCT has accomplished at the school level is offered by Forney (1989), an English teacher in a California high school requiring

BEST COPY AVAILABLE

the passage of a competency test for graduation. Forney reports that students are given a choice of seven topics (two days before the exam) in order to organize their materials and to practice writing the essay. Even though they are not allowed to bring anything with them the day of the test, the examination ends up being a memorized essay that an English teacher has already approved or that another student has written. While the exercise models the five-paragraph essay taught in preparation sessions, the students fail to write an essay showing any sort of individualized style or creativity. Problems also arise when students view passing the test not as evidence of possessing minimum competency, but as proof that their need for formal education no longer exists. Forney notes that little evidence exists showing the tests do anything towards guaranteeing that illiterate students are not continuing to pass through the system. Furthermore, no effective measures are being formulated to replace MCT, the only acceptable action possibly justifying its continued use. The only observable change noted is the placing of more and more emphasis on preparing the students for the test. Forney concludes by acknowledging the charge that high school diplomas continue to have practically no value because schools are unable to demonstrate that the holders of these diplomas even have minimum skills.

Serow (1984) further cautions that even when students initially fail a competency examination and are subsequently able to obtain a passing grade, the data do not by themselves constitute evidence of the remedial effectiveness of MCT programs. While many states point to reduced failure rates in subsequent testing attempts as evidence of remedial effectiveness, what appears to be an indication of improvement in pupils' basic skills could simply be an artifact of repeated testing. Rather than reflecting true academic increases, score gains may reflect familiarity with test items (i.e., a practice effect) or the tendency of extreme scores (in this case extreme low scores) to move toward the mean in repeated testing (i.e., regression to the mean effect).

That MCT failure rates are much higher for certain racial/ethnic minority groups than they are for White students is not surprising. The adverse impact is likely a result of prior and current discrimination, tracking, poor quality education, and other social and economic factors, over many of which neither the school nor pupil has control (Linn *et al.*, 1982). What is especially disheartening, however, is that a majority of MCT programs have chosen to emphasize the punitive aspects of testing through diploma denials, rather than maximizing their potential as diagnostic and remedial tools. The use of mandatory diploma sanctions by states having relatively large minority and low-income populations (e.g., California, New York, North Carolina, Virginia, Florida) simply confirms and perpetuates existing inequities by providing minority students with yet another educational failure (Serow, 1984).

It is not difficult to see stigmatizing tendencies within MCT — that is, the setting of minimum standards without also assuming the social responsibility for helping those who fall below them is a concern. As noted by Cohen and Haney (1980) Florida's MCT scheme, for example, gave little if any attention to the issue of remediation before the startling finding in 1977 that 40 to 50 per cent of the students failed portions of the test. MCT has also gained momentum at a time when educational funds are scarce and in light of scant evidence that it is successful in achieving any positive aims. Not only are the MCT instruments of

doubtful quality, but at the moment they merely serve to identify and not remedy the failures they define (Cohen and Haney, 1980).

Given that Chicano and other minority students are the ones most affected by MCT programs, it is the responsibility of the research and policy communities to continue to raise and seek resolutions to a number of issues concerning the use of these tests. While they initially were adopted with high levels of uncertainty, we now know that MCT has delivered little reform improvements in exchange for the severe blow it has dealt students, particularly those of racial/ethnic minority background. As noted by Airasian (1987), Ellwein *et al.* (1988) and stressed throughout this chapter, the crucial issues of testing are not only technical but also involve the allocation of privileges and opportunities. If MCT instruments are to continue in use, evidence must be shown of their effectiveness in raising standards and providing equitable learning opportunities for all children.

In conclusion, MCT is a topic of serious debate. As a major arm of the accountability movement, MCT continues to grow — but not without controversy. As we have discussed, there are far more weaknesses with MCT than there are strengths. Taking all criticisms together, it is quite clear that MCT constitutes an abuse of tests with respect to a substantial number of Chicano students. On the broad issue of MCT abuse, we agree with the issues Jensen (1980) raises:

> Although the results of MCT undoubtedly highlight a serious educational problem, I cannot see MCT as in any way contributing to the solution of the problem. It appears to me to be an unnecessary stigmatizing practice, with absolutely no redeeming benefits to individual pupils or to society. I say this not because I do not believe that individual differences in scholastic attainments cannot be reliably measured, but because I seen no utility whatsoever in drawing an arbitrary, imaginary line between 'minimal competence' and 'incompetence'. 'Competence' is an entirely relative concept. What is competence for one purpose may be incompetence for another. There can be no single all-purpose demarcation between 'competence' and 'incompetence'. *The notion is psychometrically nonsensical* [italics added]. . . . So who would possibly benefit from the extremely costly and occupationally and socially stigmatizing minimal competency testing of all graduating high school pupils? MCT is surely one of the most futile proposals to come along in public education in a decade. . . . The role of standardized tests . . . is to monitor pupil achievement periodically so as to assure its fullest development, to whatever level that might be for a given individual. *It is an abuse of tests* [italics added] to use them to assign general labels of 'competent' or 'incompetent' (pp. 724–5).

MCT is not the only form of competency testing that adversely effects Chicanos. There is also teacher competency testing, a subject we turn to next.

Teacher Competency Testing Within the larger crisis faced by the majority of Chicano students, there is brewing a smaller, but critical situation — the low and falling proportion of Chicano teachers. To best understand this grave problem,

it is helpful to place this 'crisis within a crisis' situation in a broader perspective of the minority schooling experience. In 1980, the total racial/ethnic minority elementary and secondary public school enrollment nationally was 27 per cent (Orfield, 1988). By the year 2000, the combined minority kindergarten through twelfth grade (K-12) enrollment is predicted to be 33 per cent of the total, national public school population (Smith, 1987) — a growth of 22 per cent in two decades. During the same time period, the total racial/ethnic minority teaching force in grades K-12 is projected to *decline* by 60 per cent — from 12.5 per cent in 1980 to less than 5 per cent in the year 2000 (Smith, 1987). In this section we discuss two issues pertinent to teacher competency testing and Chicanos: the Chicano teacher shortage, and technical aspects of teacher tests.

Teacher Testing and the Chicano Teacher Shortage

Orum (1986) has identified three significant factors that contribute to the small and declining percentage of the Chicano and other Latino K-12 teaching force. These influences are: (a) Latinos' low and declining college-going rate, (b) their declining preference for choosing and pursuing careers in teaching, and (c) the very high failure rate of Latinos on state-required, standardized teacher competency tests. In one of the most sustained analyses to date of the factors related to competency testing and Latino access to the teaching profession, we have identified the teacher competency test as the major obstacle in Latino teacher production (Valencia and Aburto, in press, a, b). The evidence is unequivocal: the Chicano failure rate on teacher competency tests is considerably higher compared to their White peers. For example, in 1986–87 the failure rate on the California Basic Educational Skills Test (CBEST) was 41 per cent for Chicanos (compared to only 19 per cent for White examinees (Smith, 1987). In Texas, in the period from March 1984 to June 1987, the majority (53 per cent) of Chicano students who desired to enroll in teacher education programs failed the admissions test (Pre-Professional Skills Test, PPST). The White failure rate was quite lower at 19 per cent (Smith, 1987). Based on some very recent data, the high failure rate of Chicanos on teacher tests continues unabated. In Texas, for example, nearly 3,000 teacher education program candidates took the Texas Academic Skills Program test (TASP, a recent replacement for the PPST) in September, 1989. The failure rates for Chicanos and Whites were 39 per cent and 14 per cent, respectively (Garcia, 1989).

An immediate consequence of the high fail rates of Chicano examinees on teacher tests can be seen in the growing Chicano student/Chicano teacher disparity. For example, in the mammoth Los Angeles City Unified School District, a few years ago Latino students comprised one of every two K-12 students, yet only one in ten teachers were Latinos (Crawford, 1987). On a state-wide analysis of California, Latinos fared no better. In the 1987–88 school year, Latino K-12 students were 30 per cent of the total public school enrollment in the state, but only 7 per cent of the K-12 teachers were Latino (Watson, 1988) — a Latino student/Latino teacher disparity of 77 per cent (that is, Latino teachers were underrepresented by 77 per cent). This disparity figure, by the way, is very close to the national disparity percentage (75 per cent) for Latino teachers (Valencia and Aburto, in press, a).

The predominant view is that the growing shortage of Chicano and other minority teachers contribute negatively to the education of all students in a pluralistic society (e.g., Bass de Martinez, 1988; Nava, 1985). An additional

concern is the absence of teacher role models for minority youngsters, and all what that entails — for example, passing on cultural heritage, instilling minority pride, promotion of racial understanding among all students, and so on (see Valencia and Aburto, in press, a, for an extended rationale for the value of having Latino teachers).

An especially serious consequence of the Chicano teacher shortage is the severe and worsening shortage of bilingual/multicultural teachers and the resultant impact in meeting the needs of Chicano students who are limited- or non-English-speaking. For instance, in California (the state with the largest Chicano student population) it is predicted that in the year 2000, there will be about 18,000 bilingual teachers. Yet, the actual demand to meet the needs of the thousands and thousands of linguistic minority students will be about 30,600 bilingual teachers — a projected shortfall of about 12,000 teachers (Olsen, 1988). The bilingual teacher shortage is also acute in Texas, the state with the second largest Chicano student enrollment. In 1985, Nava had this to say about the severe bilingual teacher shortage in Texas with respect to meeting the needs of over 600,000 language minority students (99 per cent of them Latino, predominantly Chicano): 'An additional 20,000 bilingually certified/endorsed teachers are needed to provide adequate equal education opportunities for these linguistically and culturally different children' (p. 34). Summing matters up, Valencia and Aburto (in press, a) observe,

> . . . the high failure rate of Latinos on teacher tests will continue to be a major contributing factor in blocking access to teaching careers in bilingual education — unless this obstacle to access is vigorously dealt with. In any event, the negative effects of teacher testing on the Latino community are here and now. (p. 21)

Technical Aspects of Teacher Tests

What is particularly distressing about the low and falling proportion of Chicanos and other minorities in the teaching profession is that the main barrier — teacher competency tests — are very questionable in how they are constructed and in what they purportedly predict. Valencia and Aburto's (in press, b) analysis of these issues center on two concerns: (a) the reliability and validity of existing paper-and-pencil teacher competency tests, and (b) the decision-making aspects of teacher testing — standard setting, that is how a predetermined cut score of a particular test is developed and used to decide who passes and fails.

Regarding reliability and validity of these tests, the existing psychometric evidence is weak and irrelevant. Also, research on the question of potential racial/ethnic bias on teacher tests is sorely needed. Based on their evaluation of the available literature and pertinent reviews, Valencia and Aburto (in press, b) come to this conclusion:

> . . . while psychometric evidence exists on current certification examinations, none of it is strong. In the case of reliability, while internal consistency estimates are high, cut score reliability has not been thoroughly examined. Criterion-related validity is extremely weak and, as has been pointed out by Haertel (1988), much of the content validity evidence is irrelevant to the uses made of licensure tests. Although

> current item bias studies and sensitivity review panels contribute to removing bias from existing examinations, no test bias research (defined as differential criterion validity for different groups) exists. (p. 21)

With respect to the cut score in standard setting on teacher tests, it has become the linchpin of the decision-making process along technical, political, and equality lines. Valencia and Aburto (in press, b) note that the cut score has two sides — one strong, one weak. One side represents omnipotence in deciding who passes and who fails a particular teacher competency test. The other side, however, is brittle and open to charges that the cut score standard cannot be defended on how it is technically decided. In that standard-setting methods use a great deal of human judgment in trying to capture and measure the notion of 'competence' during setting, there is a resultant variability in accuracy. In reference to judges being asked to speculate on the competence of unknown test-takers and to give some probability statements of item accuracy, Haertel (1988) states, 'There is simply no evidence that people can perform this kind of task with accuracy' (p. 60). As such, there are major problems in the methods and steps used to develop the linchpin of teacher testing. From a measurement and technical perspective, Haertel offers this critique of the cut score determination:

> I see absolutely no basis for asserting that the judgments of individual panelists about individual items are unbiased estimates of performance for the imagined target population of minimally competent teachers. I consider attempts to derive a meaningful cutting score by aggregating panelists' judgments to be at best a meaningless misapplication of statistical theory. (p. 61)[23]

To summarize matters, teacher competency testing as a measurement tool of the accountability movement contains some serious problems. The available reliability and validity evidence for teacher tests is weak and not particularly relevant. Regarding the development and use of cut scores, there is a growing controversy in the measurement community about their technical defensibility. Finally, there is the arbitrariness of how cut scores are set by school officials (we did not cover this issue here; see Valencia and Aburto, in press, b). As Haney and Madaus (1978) contend, the point at where the cut score on basic skills-type tests is eventually set often comes about as a compromise between apparently acceptable expectations of pass and failure rates that are politically tolerable.

With the above problems in mind, it is not difficult to conclude that the sole or near sole use of teacher competency tests to determine admissions to teacher education programs or to grant state license is a practice difficult to justify. Given all the concerns raised — coupled with the high-stakes nature of teacher testing — we argue that such testing constitutes an abusive practice. There is little doubt in the high-stakes game of teacher competency testing that a substantial number of prospective Chicano teachers are clear and big losers. Within a five-year period alone, Smith's (1987) study of nineteen states documented the alarming teacher test failure of 10,142 Latinos.[24]

How will prospective Chicano and other Latino teachers fare in the 1990s and beyond? Zapata (1988) offers this assessment of the ever-growing Latino student/Latino teacher gap, 'Projections for the future are generally bleak' (p. 20).

Such pessimism, of course, is embedded in the presumption that the status quo of teacher testing will go undisturbed as we approach the twenty-first century — a belief, we contend, that has to be challenged. We agree with Smith (1988) who voices, 'For the nation to plunge headlong into the 21st century with a public school system devoid of minority teachers is unacceptable' (p. 168). At this chapter's end we offer some research and policy ideas how educational testing and assessment might be strengthened to help promote far greater access for prospective Chicano teachers.

On Educability and Chicano Students

Up to this point in our analysis, we have concentrated on two major abusive practices of educational testing with respect to Chicano students — the administration of tests that lack good technical qualities (particularly administering instruments that have not been examined for potential cultural bias), and the use of high-stakes testing (i.e., relying on a single test or score) in educational decision-making. To conclude our discussion on this note would not only be premature but misleading. It is easy to 'blame the tests', and certainly there has been a great deal of test-bashing over the last two decades. The issue is not that *all* educational testing is invalid and leads to discriminatory outcomes. The real issue is to identify those tests that are psychometrically poor, biased, and used in unfair ways. In short, the issue is documentation, not imputation. Madaus (1986) puts matters in a way that make most sense to us: '. . . the decision is inherently linked to the tool. The question that we need to consider is what line of evidence needs to be gathered to counter critics who blame poor results on the use of an invalid test' (p. 12).

Indeed, what kind of evidence needs to be gathered? Should the evidence be only in the form of statistical, quantifiable (particularly psychometrically-driven) data? Certainly, scientifically grounded evidence derived from reliability, validity, and bias investigations are essential. We also argue, however, for the examination of extrascientific evidence, that is, understanding people's perceptions of an individual's and group's presumed ability to learn and the connection of educational testing to these perceptions.

Historically, and to some degree presently, tests have been widely used by schools in helping to decide whom — and how much — to educate. But the linkage between the measurement of scholastic aptitude of school children and their presumed capacity for scholastic learning is very shaky in its theoretical underpinnings. According to the dominant view, school learning ability is influenced primarily by the child's intellectual ability. What have risen from this perspective are institutionalized different school curricular policies and practices allegedly based on the belief that students can be hierarchically arranged as 'advanced', 'average', or 'slow' learners. The belief in this 'normal' distribution of educability — or scholastic learning ability — is one of the most entrenched assumptions in education today. Such an assumption potentially carries grave implications for Chicano students. Some scholars would have us believe that educability is largely dependent on individual intellectual ability and that social, political, and economic conditions within the schools and society are largely unrelated to '. . . *why* some of our children are so much more educable than

others' (Hawkins, 1984, p. 375). So, we contend, an additional fruitful way to discuss the abuses of educational testing regarding Chicano students lay in the concept of 'educability'.

Educability is not a new notion. The belief that the quality and quantity of schooling should be dictated by a person's perceived and/or measured abilities and other background characteristics can be seen in the writings of Plato (*ca.* 348 BC) over 2,000 years ago. Plato's educational philosophy emphasized the training of an élite, that is the offspring of the guardians or statesmen who directed the policy of the commonwealth were those who had opportunities to be formally educated. On the other hand, the 'common people', that is the business and working classes who were at the bottom of the Athenian caste system, would be directed to practical and vocational jobs (Ulich, 1950).

There is also longstanding evidence that the concept of educability cut deeply into racial questions. For example, Lyons' (1975) *To Wash An Aethiop White: British Ideas About Black African Educability, 1530–1960*, gives an evolutionary account of the belief that Black Africans were intellectually inferior to White Europeans, and the influence that idea had upon British attitudes and policies toward the education of Africans during the colonization of Africa. He is able to show strong connections between British attitudes about the alledged African intellectual inferiority and the implementation of school policies directed toward simpler, less demanding education for Black Africans. These practices were based on the belief that the presumed 'deficient mental capacity' of Black Africans placed severe limitations on how much they could benefit from education. Lyons underscores the point that ideas about human mental ability and educability superseded the testing movement by centuries:

> ... attitudes about human mental ability long antedate the concept of 'IQ' ... Britons began to formulate opinions about capacity for learning or educability at least as far back as the seventeenth century. In many respects the psychological testing movement of the twentieth century represents less a startlingly new departure and more a continuation of a type of investigation which had been going on for at least three centuries. (pp. xi–xii)

In more recent history, it is interesting to note that Alfred Binet (the co-developer of the first intelligence tests) used the term 'educability' in his 1910 book, *Modern Ideas About Children*, whose first chapter was entitled 'The Educability of Intelligence'. Binet, one of the early remedial educators, was a firm believer that intelligence could be 'trained'. His classes for the mentally retarded in Paris in 1909 consisted of a curriculum that emphasized the training of memory, attention, judgment, and other factors of intelligence he believed important (Kirk, 1973).

Ironically, Binet, the father of mental testing and a strong believer in the modifiability of intelligence, held a minority opinion with respect to educability. The dominant position held by early psychologists at the turn of the century was: intelligence is fixed at the point of conception; intelligence (as measured by IQ) is constant over time; intelligence is unalterable by the environment (Kirk, 1973). How this notion of educability influenced views on the relations among intelligence, school learning, and vocational guidance was exemplified by Lewis

Terman, a dominant figure during the advent of the testing movement in the United States.

> . . . the grade of school work which a child is able to do depends chiefly upon the level of mental development he has attained . . . *The limits of a child's educability can be fairly accurately predicted in the first school year* [italics added]. By repeated tests these limits can be determined accurately enough for all practical purposes by the end of the child's fifth or sixth school year.
>
> Vocational guidance is not, and may never be, an exact science. Nevertheless, *intelligence tests will be of value even if they tell us nothing more than that reasonable success in a given vocation is or is not compatible with the general mental ability which a particular individual possesses* . . . [italics added] (Terman, 1920, p. 21)[25]

Anastasi (1984) offers some sharp insights to common misconceptions about how measured aptitude and achievement were believed to be related during the 1920s. A case in point Anastasi refers to is a popular and widely used textbook of the times — Frank N. Freeman's *Mental Tests* (1926). Anastasi comments on two unwarranted assumptions discussed in Freeman's book. First, intelligence tests gave a measure of innate capacity (i.e., not dependent on training). Second, all school achievement depended on the same, singular intellectual capacity.

Beliefs in fixed intelligence and the minimal impact of the environment on modifying intelligence were entrenched in the intelligence testing movement up to about World War II. Beginning in the 1940s and up to the present, however, mounting theory and evidence have challenged beliefs about the immutability of intelligence (Hunt, 1972). For example, the works of Hunt (1961) and Piaget (1952) represent prominent research with respect to new conceptions of intelligence and the malleability of intellectual development.

In summary, thoughts on the relation between intelligence and school learning — like the debate about the relative contributions of nature and nurture to intelligence — have hit peaks and valleys for decades. Dabney (1980) has explained the historical essence of the educability issue in this way:

> The historical emphasis upon capacity for learning has been to perceive school learning as primarily dependent upon the presumed ability of the student, rather than upon the quality of the learning environment. However, there appears to be a growing recognition that school failure and student achievement are socially determined. Even so . . . such recognition has not prevented new interpretations of these failures which blame the victims and often co-exist with arguments about innate or class deficiencies. (p. 13)

In the case of Chicanos, as we moved year by year into the nascent period of the intelligence testing movement of the twentieth century, stronger and stronger connections are seen between American attitudes about alleged Chicano intellectual inferiority and the implementation of school policies directed toward simpler, less challenging education (cf. Gonzalez, 1974a). Those policies were based, in part, on the pseudo-scientific, even racist beliefs, that the presumed deficient

mental capacity of Chicanos and other racial/ethnic minority groups placed severe limitations on how much they could benefit from schooling.

More recently, the issue of educability of Chicano students has recently received national exposure in the inspiring 1988 motion picture, *Stand and Deliver*, in which the character of East Los Angeles teacher Jaime Escalante was brought to life by the widely acclaimed performance of actor Edward James Olmos. In the beginning of their tough, intellectual odyssey, Escalante speaking slowly and unflinchingly warns his Chicano students: 'You already have two strikes against you. There are some people in this world who will assume that you know *less* than you do, because of your name and your complexion'. Later in the movie, Escalante's foreboding words materialize. Because his Advanced Placement Calculus students score extremely high on the calculus test, have very similar errors, and finish the test with plenty of time to spare, the Educational Testing Service suspects these irregularities to be evidence that Escalante's students cheated. In a very dramatic scene, Escalante confronts the two psychometricians from the ETS who are in charge of the investigation. Escalante demands to see the evidence of cheating, but the investigators refuse to accommodate him. Furiously, Escalante lashes out at the two men: 'These scores would have never been questioned if my kids did not have Spanish surnames and come from barrio schools! You know that!'

Let us now conclude by placing the notion of educability in its proper place regarding educational testing and the schooling of Chicano students. Granted, it is very difficult to pin down precisely an intangible idea as educability and its relationship to the education of Chicanos. We do speculate, however, that in light of our discussion thus far, the concept of educability as a value-laden, extrascientific notion has helped to mold historical and current thought on the nature of schooling for Chicanos. A major feature of this analysis is that the abusive practices of educational testing have been partially instrumental in creating school inequality. We contend that the ideological configuration of educability *also* needs to be taken into consideration why barriers to educational opportunities for Chicanos exist. The perspective that educational tests have served as oppressive, sorting tools in the schools has some value in theory building. This thesis, however, as currently structured is essentially mechanistic, deterministic, and simplistic in scope. It gives too much credit to tests as sole forgers of inequality while failing to understand that a great deal of oppression also lies in a fundamentally unequal society that views the educability of working-class and Chicano students as limited.

Improving Educational Testing for Chicanos

Now, let us shift gears and travel into a brighter territory. In this concluding section of the chapter, we offer a number of research and policy oriented ideas how educational testing — particularly test use — might be improved to help promote Chicano school success. These suggestions cover eight topics: test bias research, test translation, performance v. capability, multiple data sources, criterion-referenced testing, short-term solutions, science and ethics, and educability.

Test Bias Research

As we have seen, there are some individuals in the measurement community who contend that the question of cultural bias in educational testing — particularly intelligence tests — is a closed issue. On the contrary, we disagree. The subject of cultural bias with respect to Chicano students remains an open issue whose resolution lies in the years ahead.

We advocate continued research into the subject of cultural bias of various educational tests, hopefully with more sophisticated paradigms and designs that cover the complex relations among the testing situation, such as the actual item, the instructions, the examiner, and aspects of the examinee including at least attitudes toward the task, motivation, and the finer grained mental processes (e.g., attention, memory; information processing; cf. Scheuneman, 1984).

Test Translation

The number of non-English-speaking (NEP) and limited-English-speaking (LEP) Chicano students will continue to increase as we enter the 1990s, and their dramatic growth is predicted to continue as we move into the twenty-first century.[26] The psychoeducational assessment of these students presents a difficult problem and challenge for the field of school psychology (Figueroa, Sandoval and Merino, 1984). An especially troubling area is the lack of standardized intelligence and achievement tests that can be administered to NEP and LEP students.

A recent approach to this shortage issue has been the development of translated versions of extensively used, individually administered cognitive tests (see Valencia and Rankin, 1985, for a brief discussion). Although we applaud the development of such tests (including a variety of psychoeducational assessment tests), "it is critical that these tests meet the minimal psychometric properties of reliability and validity and are free of bias. Thus, intensive comparative research on these translated tests and their English equivalents is strongly encouraged" (Valencia and Rankin, p. 206).

Performance v. Capability

The educational community will have taken a big step forward if those who are in the business of undertaking the individualized psychoeducational assessment of Chicano students would understand and acknowledge the critical distinction between *performance* and *capability*. What students *do* in a testing situation (their observed *performance*) is not always congruent what they *could do* (their *capability*). Factors helping to create this discrepancy during testing are referred to as situational influences. Interindividual differences, for example, in examinees' test-taking skills, motivational level, responses to time pressures of speeded tests, and so on, can and do account for variability in test performance among students (Henderson and Valencia, 1985). For Chicano and other students from culturally/

linguistically diverse backgrounds, their familiarity with the language and cultural content of the test (i.e., the degree of cultural loading) is a particularly important factor to consider during testing. In sum, school psychologists and others who routinely administer tests should be vigilant of the potential situational influences that may lower the test performance of Chicano students.

Multiple Data Sources

'In all . . . educational decisions, test scores provide just one type of information and should always be supplemented by past records of achievement and other types of assessment data. *No major educational decision should ever be based on test scores alone*' (Gronlund, 1985, p. 480).

We daresay by now the above version of an admonishment we have been reiterating throughout should be quite familiar to the reader. We cannot repeat it enough. It should be kept in mind that testing is only one component of three types of information that can be collected during the process of psychoeducational assessment. The other two sources of diagnostic information are 'observations' and 'judgments' (Salvia and Ysseldyke, 1988). Each of the three types can be collected by a diagnostician or another person, whom Salvia and Ysseldyke refer to as 'direct information' and 'indirect information' sources, respectively.

Suffice it to say that conceptualizing psychoeducational assessment as a tripartite structure and process can greatly enhance the gathering of data, diagnosing intermdividual strengths and weaknesses based on these multiple sources, and making decisions on how to improve schooling for students. The use of multiple, informed data sources (e.g., tests; parental and teacher informants; classroom observations; medical records) have the potential to provide a rich database and also to improve the credibility of the various sources (Valencia, 1982). We strongly encourage that multi-measurement efforts be utilized in the psychoeducational assessment of Chicano students.

Criterion-Referenced Testing

Although the notion of an absolute versus a relative standard of testing can be traced to about 1913, research investigations and practical applications of criterion-referenced testing did not truly begin until the early 1960s (Berk, 1984). In the simplest sense, such tests '. . . are designed expressly for interpreting an individual's performance in terms of what he or she can and cannot do irrespective of the performance of other students . . . ' (Berk, 1984, p. 1). Or, a more precise definition: 'A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards' (Glaser and Nitko, 1971, p. 653; cited in Nitko, 1984).

The use of tests to identify learning problems and to utilize such results to modify instruction of individual students is a very commendable educational practice. Furthermore, it is a form of testing that the schools should use more of (particularly teacher-made tests). The payoffs for school learning appear to be quite substantial. We share Nitko's (1984) call for research in criterion-referenced

testing, as it seems important for the advancement of both instructional theory and practice. With respect to Chicano students, given their typically low academic performance, well-developed criterion-referenced tests that provide useful diagnostic feedback could be a bonanza in improving these students' academic progress.[27]

Short-Term Solutions

Until the time comes when broad-based, workable, nondiscriminatory testing and assessment *vis-à-vis* Chicano students materialize, it is necessary to identify and implement short-term solutions to the abuses of educational testing. As a case in point, let us take teacher competency testing.

In a recent analysis, we discuss numerous practical strategies of test reform and other means to improve access for Chicano and other Latino students who aspire to become teachers (Valencia and Aburto, in press, b). Examples include the modifing (i.e., lowering) of cut scores of existing paper-pencil teacher tests, implementing multiple data sources of assessment (including performance-based assessment), and using numerous ways to identify, recruit, test diagnostically, and offer remediation for perspective Latino teachers. Taken together, some of these suggestions require monetary commitments, others mean institutional-wide commitment, while others require convincing policymakers that such ideas will work.

Science and Ethics

We have argued repeatedly that tests do not exist in a vacuum. They have social consequences. As such, there needs to be a striving for a unified view of test validity that integrates both the *science* and the *ethics* of assessment. In a recent treatise, Messick (1989) presented a very thoughtful paper on the importance and necessity for the integration of science and ethics. In brief, Messick contends that test validity and values are one imperative, not two. Thus, test validation implicates both science and ethics. This unified conceptualization of validity, according to Messick, integrates both the scientific and ethical underpinnings of how tests are interpreted and used. The following, we believe, gets to the fundamental nature of this inherent tie between meaning and values in test validation:

> . . . it is simply not the case that values are being added to validity in the unified view. Rather, values are intrinsic to the meaning and outcomes of the testing . . . This makes explicit what has been latent all along, namely, that validity judgments *are* value judgments. (Messick, 1989, p. 10)

It is easy to see that in the case of Chicanos — and other groups who have at times been victimized by abusive testing practices — such a unified view of test validity, if universally accepted, would certainly help to promote nondiscriminatory assessment.

Educability

Our list of ideas for improving educational testing for Chicanos ends with a reference to a previously discussed notion — educability. The main point posited here is difficult to attain, but relatively simple to grasp. Those in the educational system (e.g., teachers, school psychologists, administrators, elected officials) who have direct and indirect contact with Chicano students must come to grips with any preconceived, negative notions they hold of the educability of Chicanos. It would be wonderful if all educators (especially kindergarten through twelfth-grade teachers) would embrace the proposition that *every Chicano student has an infinite capacity to learn*. If this fundamental premise becomes widely accepted in the schools, then it would not be so easy to explain away low test performance as simply due to a perceived learning problem inside the Chicano student or to an alledgedly impoverished home environment.

Conclusions

We leave the reader with two final observations. First, as we make the transition into a new decade and move slowly into the twenty-first century, it is likely that educational testing will increase in frequency and in its variants. In particular, the accountability movement with its emphasis on measuring 'competence' will continue to be driven by an 'If it moves, test it!' mentality. If 'the more testing the better' philosophy of the 1970s and 1980s remains unchallenged in the 1990s, then such an ideology is likely to become more deeply entrenched. On the optimistic side, a serious challenge to the status quo in educational testing can be mounted. This reform effort will likely include a number of features we have discussed throughout this analysis — for example, the development of test instruments that are nonbiased, the use of multiple data sources of assessment, and a unified conceptualization of test validity. In that ' . . . testing of abilities has always been *intended* as an impartial way to perform a political function — that of determining who gets what' (Cronbach, 1984, p. 5), we urge those who are interested in Chicano students to become vigilant and active in bringing about the needed shifts in educational testing reform in the very near future.

Our second and last observation flows from the first one. Reform in educational testing with respect to Chicanos needs to be placed in the broader subject of school reform. To a very large extent, the typically low performance of Chicano students on norm-referenced achievement tests, on MCT, on school-based competency tests, and so on, is just one manifestation of the poor schooling they receive. We argue that given the massive schooling problems Chicanos face (e.g., school segregation; curriculum differentiation; disparities in school financing), their low test performance is not surprising.

To understand this linkage between low test performance and schooling inequalities, it is helpful to mention 'opportunity to learn', a construct which is receiving some attention in the measurement community (Tittle, 1982). Basically, the notion of opportunity to learn deals with the fit — or lack of fit — between the content of a test (i.e., those samples of behavior that are measured), and the formal curriculum (i.e., that which is taught and learned in school). The implication for the testing of Chicano students is clear: if Chicanos are not given

the opportunity to learn the test material on which they later will be tested, then it is not surprising that their test scores will be low. As such, there are increasing instances in which claims as the following are being voiced: 'Testing children on what they have not been taught and then stigmatizing their "failure to learn" is a fundamental form of discrimination' (Hanson, Schutz, and Bailey, 1980, p. 21).

In the final analysis, it is refreshing to see a renewed interest in educational issues regarding Chicano students. Let us hope that as we begin the short trek to the next century there will be more light than heat. The abusive testing practices that we have described and discussed here can be remedied with better scientific work and ethical judgments. As reform in educational testing takes shape and evolves, it is vital to link it inextricably with broad-based school reform, lest such testing reform efforts become mere palliatives. The many issues we have outlined demand to be answered and present the research, educational, measurement, and policymaking communities with significant challenges in the years ahead.

Notes

1 The focus is on standardized (norm-referenced) tests. At the chapter's end, there will be a brief discussion of criterion-referenced tests.

2 Standardized, group-administered 'achievement' tests form a third major category of educational tests. Such tests are '. . . designed to measure the amount of knowledge and/or skill a person has acquired usually as a result of classroom instruction . . .' (Lyman, 1986, p. 158). By far, achievement tests are the most widely administered tests in the nation's schools. It is typical for school districts to administer such tests on a routine basis (usually during the spring). The purposes of standardized, group-administered tests include the following: to determine a pupil's developmental level so instruction can be adapted to individual needs, to diagnose a pupil's strengths and weaknesses, and to provide meaningful data that can be reported to parents (Wiersma and Jurs, 1985).

Achievement tests, as described above, have been at the center of controversy *vis-à-vis* Chicano and other minority students for some time. A frequent charge is that the results from such tests are used, in part, to sort students into ability groups for instructional purposes (e.g., see Oakes, 1985). In that Chicano students typically score lower on achievement tests, it is claimed that these students often end up in the 'low-ability' groups or tracks and thus receive inferior schooling. Although there are data to support the contention that Chicano students are disproportionately overrepresented in low-ability groups and underrepresented in high-ability groups throughout schools in the southwestern United States (e.g., US Commission on Civil Rights, 1974), it is difficult to ascertain the direct link between test performance and grouping practices. That is, there are few hard data showing that teachers *actually use* the scores from standardized, group-administered achievement tests as a major mechanism in deciding who is placed in which instructional groups. In short, there is no argument that ability grouping at the elementary school level and tracking at the secondary level exist. The issue is, how are achievement test scores used (if indeed they are used) to help shape the teacher's decision-making when it comes to grouping? In the absence of empirical findings, we have elected not to focus on the potential abuses of achievement tests.

3 In some cases these functions or purposes are discussed in the context of 'assessment', that is the process of collecting test information as well as other forms of data (see, for example, Salvia and Ysseldyke, 1988).

4 Resnick's (1979) framework centers on IQ testing. Yet, we think the functions discussed by Resnick can be generalized, in part, to other types of tests (i.e., achievement and competency).

5 In that group-administered intelligence tests have been for the most part banned throughout the nation, their 'practical function in schooling' is confined to psychoeducational assessment in special education. Yet, one can argue (as does Resnick) that intelligence tests historically have played, and contemporarily play, symbolic roles. The present authors contend that most — if not all — of educational testing also has symbolic roles, as well as practical purposes as earlier described.

6 There is, however, a recent publication that challenges this thesis. In a provocative article, Rafferty (1988) contends that educational historians have greatly exaggerated the role intelligence testing played in ability grouping placements of Chicano and other minority students in the 1920s and 1930s in Los Angeles. Rafferty's case study does shed some light on the role of IQ testing — particularly how classroom teachers may or may have not used test results for grouping purposes. We think, however, that her thesis and final analysis are misleading. Rafferty criticizes other educational historians for stating that IQ tests were used *exclusively* for educational placement. Our reading of the same material (e.g., Gonzalez, 1974a) leads us to believe that such historians *did* speak to *other* discriminatory factors than just IQ testing.

7 In the early years of intelligence testing, Chicano students — as a whole — typically performed about 10–20 IQ points lower than their Anglo peers (Valencia, 1980).

8 It is important to add that there is evidence indicating a marriage between the local industry (in this case, Los Angeles) and the school. Gonzalez (1974b) states '. . . the vocational courses were incorporated into the curriculum on the basis of labor needs of business and industry. . . . Counselors were the link between the requirements of industry and the school' (p. 299, also see Gonzalez, 1974a). These self-serving connections between business and education are good examples, we believe, of the negative social consequences that intelligence testing indirectly helped promote for Chicanos.

9 Gonzalez (1974b) reports that by 1929, there were 2,500 children enrolled in eleven development centers in Los Angeles schools. Ten of the centers were in 'laboring class communities', and of the total enrollment, Chicano children were '. . . highly represented in five of them, constituting the entire population of one, one-third in two, and one-fourth in two others' (p. 298). The development centers were geared to training unskilled and semi-skilled workers for industry (e.g., menial occupations in restaurants, laundries, and agriculture). Gonzalez offers an interesting insight to the development center business connection by quoting a school administrator who viewed the centers as bonanzas to local industry:

> . . . several employers have told us that a dull girl makes a very much better operator on a mangle than a normal girl. The job is purely routine and is irksome to persons of average intelligence, while the sub-normal seems to get actual satisfaction out of such a task. Fitting the person to the job reduces the 'turn over' in an industry and is, of course, desirable from an economic point of view. (p. 298)

10 For valuable insights to the waxing and waning of intelligence and other forms of testing over the decades, see Cronbach (1975) and Haney (1981).

11 Cultural bias is just one form of 'bias', a general term. Psychometric bias can also occur along other group memberships — e.g., social class, sex, and age.

12 The dichotomized framework of test bias and test unfairness is not without critics. For example, Hilliard (1984) states: 'The important demonstration that is needed is not an empirical demonstration of test bias, but of test utility. None of the discussion over the presence or absence of bias should be allowed to obscure that central matter' (p. 166).

13 Valencia (1988) reviewed psychometric research with respect to the McCarthy Scales of Children's Abilities (McCarthy, 1972) and Latino children (predominantly Chicano). The investigations reviewed were primarily single population studies. In such reviews, statements are drawn whether or not the observed validity and/or reliability coefficients in the individual studies are of acceptable magnitudes to conclude that the instrument has sound psychometric properties. For example, Valencia (1988) concluded:

> In the case of Puerto Rican children, the psychometric knowledge base is simply too sparse to allow any recommendation about the McCarthy's use as an assessment tool . . . For Mexican American children, our knowledge of the McCarthy's psychometric properties is considerable. Recent research has produced a good variety of studies and a solid base of knowledge from which to build. For English-speaking Mexican American children, particularly preschoolers, the McCarthy generally appears to be psychometrically sound. Based on the available evidence, it appears justified to recommend the McCarthy as an instrument in the psychoeducational assessment of these children. It is admonished that performance on the Verbal Scale, however, be interpreted with extreme caution for reasons cited earlier. (p. 100)

14 In a number of these cross-racial/ethnic studies, other minorities (e.g., Black and American Indian children) have also been subjects of WISC–R research (e.g., Oakland and Feigenbaum, 1979; Reschly and Sabers, 1979). For a review of test bias research involving Chicanos and Whites who were administered intelligence measures other than WISC–R, see Jensen (1980) and Valencia (1990).

15 In such studies, the criterion variable is typically a standardized measure of academic achievement.

16 We underscore the terms 'English-speaking' and 'native-born' as this is the context in which most conclusions about nonbiased tests are made. For example, Jensen (1980) when bringing *Bias in Mental Testing* to an end, places this caveat about the existence of bias-free mental tests:

> These conclusions are confined to native-born subpopulations within the United States, not because of any evidence on immigrant groups or on populations outside the United States that is at odds with the present conclusion, but only because of the lack of relevant studies that would warrant broader conclusion. (p. 715)

17 In the content (item) validity study (Valencia and Rankin, 1988), only very negligible amounts of bias against Chicanos were found on the Mental Processing Scale. On the other hand, pervasive item bias was observed on the Achievement Scale.

18 Incidentally, with respect to K–ABC bias research involving Chicanos, no investigations other than those by Valencia and Rankin could be located.

19 However, it is important to comment, as does Mitchell (1984) that the test publishing business is an odd mixture of positive and negative features. On one hand, there are the major, large test companies, some which employ highly capable measurement specialists. Conversely, there are the numerous publishers in the 'cottage industry' of test publishing where quality control is typically lacking.

20 In some states (e.g., Texas) variants of the school-based competency testing can also serve as MCT used for diploma award or denial (see Graves, 1989).

21 It is also common for state school officials to report test results that compare scores *across* districts. For example, in 1989 a news article in the *San Jose Mercury News* (Watson, 1989) compared CAP scores of twenty-two school districts in four counties in the peninsula location of the Bay area (California). Another example — in Texas a

1989 report by the Texas Education Agency compared TEAMS test scores for the eight largest urban school districts in the state (Effective Schools Data Resource Unit, 1989).

22 For example, in the CAPs in California,

> Schools throughout the state are ranked by two methods: one compares them to each other on a scale of one to 99. The other organizes schools into groups based on students' background information including socioeconomic level, percent of students with limited English-language ability, student mobility and percent of students from families receiving Aid to Families with Dependent Children. (Watson and Kramer, 1989, p. 28)

It is well known that some of the variables (e.g., socioeconomic level) in the second method above (sometimes called the 'band' comparison) are often proxies for racial/ethnic backgrounds of students. In short, in light of these variables — coupled with the high degree of racial/ethnic segregation in California's schools — it is typical in the band comparison that a working-class Chicano district (or school) be compared to another 'similar' district (or school). Likewise, more affluent White districts are compared to other 'similar' districts. In the case of Chicano working-class districts, such CAP percentile rank comparisons between similar districts tend to show relatively good rankings because comparisons are made only between them and not relative to higher socioeconomic, White districts. If Chicano working-class districts were to be compared with more affluent White districts, then the *real* status (i.e., low-achieving levels) of the former would be revealed.

We have no quarrel with CAP fulfilling its public accountability function. We do, though, find the band comparison indefensible. In our opinion, CAP comparisons as such may actually be harmful to Chicano and other working-class students enrolled in predominantly minority schools. That is, by using this method of comparison, school districts may be creating false impressions of satisfactory progress for these students and their parents. Compared to students in White, middle-class and more affluent districts, working-class minority students are — in reality — performing academically at very low levels. Perhaps this illusory sense of satisfactory academic achievement helps to shape perceptions among students, parents, and educators that schooling is fine and the status quo need not be challenged. Although speculative on our part, it is likely that school-based competency testing programs that use the band comparison may become nothing more than substitutes for the educational barriers erected by past abuses in group-administered intelligence tests. (Finally, for a measurement critique of the band comparison, see Cronbach, 1984.)

23 In addition to this limitation of technical indefensibility, there are two additional criticisms of standard setting: the nonexistent evidence for the establishment of a decision rule with respect to classification errors, and the arbitrariness — or political nature — of cut score determination. In that space does not permit an overview of these two limitations, see Valencia and Aburto (in press, b) for a discussion.

24 Those candidates failing were individuals attempting to enter teacher education programs or trying to obtain a teaching credential. Smith's survey documented nearly 38,000 members of minority groups failing teacher tests. In addition to the over 10,000 Latinos who failed during the five-year period, there were: 21,515 Blacks, 1,626 Asian Americans, 716 American Indians, and 3,718 members of other minority groups. Smith admonishes that the data are likely underestimations in that some states do not disaggregate test results by racial/ethnic group and because some states were not able to provide data for all test administrations.

25 Terman's thoughts about the predictive nature of IQ indicated a rather tight fit between measured intelligence of a child and the subsequent, attained occupational status during adulthood. Note the specificity of the following:

> Preliminary investigations indicate that an IQ below 70 rarely permits anything better than unskilled labor; that the range from 70 to 80 is preeminently that of semi-skilled labor, from 80 to 100 that of the skilled or ordinary clerical worker, from 100 to 110 or 115 that of the semi-professional pursuits; and that above all these are the grades of intelligence which permits one to enter the professions or the larger fields of business. Intelligence tests can tell us whether a child's native ability corresponds approximately to the median for: (1) the professional classes; (2) those in the semi-professional classes; (3) ordinary skilled workers; (4) the semi-skilled laborers; or (5) unskilled laborers; and *this information is of great value in planning a child's education* [italics added]. (Terman, 1920, p. 31)

26 Pallas *et al.* (1988) discuss long-term projections for the growth of children who speak a primary language other than English. In 1982, there were just under 2 million of such children. The number of NEP/LEP children is projected to triple (to about 6 million) by 2020.

27 We also highly encourage further research and classroom utilization of the principles of mastery learning theory. In brief, this instructional strategy uses tests in a monitoring fashion to guide closely a student's learning. All pupils are expected to have high mastery of material in all course objectives. In regular classroom instruction, time allotted for learning is generally held constant and classroom achievement is expected to have a wide variance. In mastery learning instruction, however, time for learning becomes variable and achievement is expected to have a restricted range (i.e., little variance, high performance) (see Gronlund, 1985).

References

Aguirre, A., Jr. (1979a) 'Chicanos, intelligence testing, and the quality of life', *Educational Research Quarterly*, **4**, pp. 3–12.

Aguirre, A., Jr. (1979b) 'Intelligence testing and Chicanos: A quality of life issue', *Social Problems*, **27**, pp. 186–95.

Airasian, P.W. (1987) 'State-mandated testing and educational reform: Context and consequences', *American Journal of Education*, **95**, pp. 393–412.

Airasian, P.W. (1988) 'Symbolic validation: The case of state-mandated, high stakes testing', *Educational Evaluation and Policy Analysis*, **10**, pp. 301–13.

Alley, G. and Foster, C. (1978) 'Nondiscriminatory testing of minority and exceptional children', *Focus on Exceptional Children*, **9**, pp. 1–14.

Anastasi, A. (1984) 'Aptitude and achievement tests: The curious case of the indestructible strawperson', in B.S. Plake (Ed.) *Social and Technical Issues in Testing: Implications for Test Construction and Usage*, Hillsdale, NJ, Erlbaum, pp. 129–40.

Archer, E.L. and Dresden, J.H. (1987) 'A new kind of dropout: The effect of minimum competency testing on high school graduation in Texas', *Education and Urban Society*, **19**, pp. 269–79.

Baratz, J.C. (1980) 'Policy implications of minimum competency testing', in R.M. Jaeger and C.K. Tittle (Eds) *Minimum Competency Achievement Testing: Motives, Models, Measures and Consequences*, Berkeley, CA, McCutchan, pp. 49–68.

Bass de Martinez, B. (1988) 'Political and reform agendas' impact on the supply of Black teachers', *Journal of Teacher Education*, **39**, pp. 10–13.

Berk, R.A. (Ed.) (1982) *Handbook of Methods for Detecting Test Bias*, Baltimore, MD, Johns Hopkins University Press.

Berk, R.A. (Ed.) (1984) *A Guide to Criterion-referenced Test Construction*, Baltimore, MD, Johns Hopkins University Press.

Bersoff, D.N. (1984) 'Social and legal influences on test development and usage', in B.S.

Plake (Ed.) *Social and Technical Issues in Testing: Implications for Test Construction and Usage*, Hillsdale, NJ, Erlbaum, pp. 87–109.

Bigelow, R.A. (1982) [Review of *Bias in Mental Testing*] *Educational Researcher*, **11**, pp. 21–3.

Binet, A. (1910) *Modern Ideas about Children*, Paris, E. Flamarion.

Blum, J. (1978) *Pseudoscience and Mental Ability: The Origins and Fallacies of the IQ Controversy*, New York, Monthly Review Press.

Bowles, S. and Gintis, H. (1976) *Schooling in Capitalist America: Education Reform and the Contradictions of Economic Life*, New York, Basic Books.

Boykin, A.W. (1986) 'The triple quandary and the schooling of Afro-American children', in U. Neisser (Ed.) *The School Achievement of Minority Children: New Perspectives*, Hillsdale, NJ, Erlbaum, pp. 57–92.

Brown, P.R. and Haycock, K. (1985) *Excellence for Whom?* Oakland, CA, The Achievement Council.

Buros, O.K. (1974) *Tests in Print II*, Highland Park, NJ, Gryphon Press.

Buros, O.K. (1978) *The Eighth Mental Measurements Yearbook*, Highland Park, NJ, Gryphon Press.

Callahan, R.E. (1962) *Education and the Cult of Efficiency: A Study of the Social Forces that have Shaped the Administration of the Public Schools*, Chicago, IL, University of Chicago Press.

Christner, C. and Molde, L.H. (1988–89) *Priority Schools: The Second Year*, Austin, TX, Department of Management Information, Office of Research and Evaluation, Austin Independent School District.

Cleary, T.A. (1968) 'Test bias: Prediction of grades of Negro and white students in integrated colleges', *Journal of Educational Measurement*, **5**, pp. 115–124.

Cohen, D.K. and Haney, W. (1980) 'Minimums, competency testing, and social policy', in R.M. Jaeger and C.K. Tittle (Eds) *Minimum Competency Achievement Testing: Motives, Models, Measures and Consequences*, Berkeley, CA, McCutchan.

Crawford, J. (1987) 'Bilingual education: Language, learning, and politics', *Education Week*, **6**, pp. 19–50.

Cronbach, L.J. (1975) 'Five decades of public controversy over mental testing', *American Psychologist*, **30**, pp. 1–14.

Cronbach, L.J. (1984) *Essentials of Psychological Testing*, 4th ed., New York, Harper and Row.

Darney, M.G. (1980) 'The Gifted Black Adolescent: Focus upon the Creative Positives', paper presented at the Annual International Convention of the Council for Exceptional Children, Philadelphia, PA, April (ERIC Document Reproduction Service No. ED 189 767).

Dean, R.S. (1980) 'Factor structure of the WISC-R with Anglos and Mexican-Americans', *Journal of School Psychology*, **18**, pp. 234–9.

Deutsch, M., Fishman, J., Kogan, L., North, R. and Whiteman, M. (1964) 'Guidelines for testing minority group children', *The Journal of Social Issues*, **20**, pp. 127–45.

Diana v. Board of Education (1970) Civil action no. C-70-37 (N.D. Cal.).

Dunn, L.M. (1987) *Bilingual Hispanic Children on the US Mainland: A Review of Research on their Cognitive, Linguistic, and Scholastic Development*, Circle Pines, MN, American Guidance Service.

Effective Schools Data Resource Unit (1989) *Executive Summary for the Analysis of Texas Educational Assessment of Minimum Skills Test Results from the 1987–88 Test Cycle: A Comparison of San Antonio Independent School District with the State and Seven Urban School Districts*, Austin, TX, Texas Education Agency, Division of Accreditation.

Eisenburg, T.E. and Rudner, L.M. (1988) 'State testing of teachers: A summary', *Journal of Teacher Education*, **39**, pp. 21–2.

Ellwein, C.M., Glass, G.V. and Smith, M.L. (1988) 'Standard of competence: Propositions on the nature of testing reforms', *Educational Researcher*, **17**, pp. 4–9.

BEST COPY AVAILABL

EYSENCK, H.J. and KAMIN, L. (1981) *The Intelligence Controversy*, New York, Wiley.

FASS, P.S. (1980) 'The IQ: A cultural and historical framework', *American Journal of Education*, **88**, pp. 431–58.

FEDERAL REGISTER (1977) 'Education of Handicapped Children', Regulations Implementing Education of All Handicapped Children Act of 1975, August, pp. 42474–518.

FIGUEROA, R.A., SANDOVAL, J. and MERINO, B. (1984) 'School psychology and limited-English-proficiency (LEP) children: New competencies', *Journal of School Psychology*, **22**, pp. 131–43.

FLYNN, J.R. (1980) *IQ and Jensen*, London, Routledge and Kegan Paul.

FORNEY, W. (1989) 'Striving for mediocrity? When MCTS become the maximum', *Thrust*, **18**, pp. 31–3.

FREEMAN, F.N. (1926) *Mental Tests: Their History, Principles, and Applications*, Boston, Houghton Mifflin.

GARCIA, G.X. (1989) '74% of graduates pass TASP tests: Failure rate among Black students 52%', *Austin American-Statesman*, November 25, pp. A1, A8.

GLASER, R. and NITKO, A.J. (1971) 'Measurement in learning and instruction', in R.L. THORNDIKE (Ed.) *Educational Measurement*, 2nd ed., Washington, DC, American Council on Education, pp. 625–70.

GONZALEZ, G.C. (1974a) 'The System of Public Education and Its Function within the Chicano Communities, 1910–1930', unpublished doctoral dissertation, University of California, Los Angeles.

GONZALEZ, G.C. (1974b) 'Racism, education, and the Mexican community in Los Angeles, 1920–1930', *Societas*, **4**, pp. 287–301.

GOULD, J. (1980) 'Jensen's last stand', *New York Review of Books*, May 1.

GRAVES, D. (1989) '10,000 students fail TEAMS twice', *Austin American-Statesman*, November 21, pp. B1, B4.

GRAVES, D. and BREAUX, B.J. (1989) '18 Austin schools receive low rating from state agency: District trustee plans to review results of minimum skills tests taken last year', *Austin American-Statesman*, September 6, p. B2.

GREEN, D.R. (1975) *What Does It Mean to Say a Test is Biased?* paper presented at the meeting of the American Educational Research Association, Washington, DC, March/April (ERIC Document Reproduction Service No. ED 106 348).

GRONLUND, N.E. (1985) *Measurement and Evaluation in Teaching*, 5th ed., New York, Macmillan.

Gutkin, T.B. and REYNOLDS, C.R. (1980) 'Factorial similarity of the WISC-R for Anglos and Chicanos referred for psychological services', *Journal of School Psychology*, **18**, pp. 34–9.

HAERTEL, E.H. (1988) 'Validity of Teacher Licensure and Teacher Education Admissions Tests', paper prepared for the National Education Association and Council of Chief State School Officers.

HANEY, W. (1981) 'Validity, vaudeville, and values: A short history of social concerns over standardized testing', *American Psychologist*, **36**, pp. 1021–34.

HANEY, W. and MADAUS, G. (1978) 'Making sense of the competency test movement', *Harvard Educational Review*, **48**, pp. 462–84.

HANSON, R.A., SCHUTZ, R.E. and BAILEY, J.D. (1980) *What Makes Achievement Tick: Investigation of Alternative Instrumentation for Instructional Program Evaluation*, Los Alamitos, CA, Southwest Regional Laboratory for Educational Research and Development.

HAWKINS, T. (1984) 'Vote of confidence. Commentary in "Backtalk"', *Phi Delta Kappan*, **65**, p. 375.

HENDERSON, R.W. and VALENCIA, R.R. (1985) 'Nondiscriminatory school psychological services: Beyond nonbiased assessment', in J.R. BERGAN (Ed.) *School Psychology in Contemporary Society*, Columbus, OH, Charles E. Merrill, pp. 340–77.

HENDRICK, I. (1977) *The Education of Non-whites in California, 1849–1979*, San Francisco, R & E Associates.

HERRON, M.D. (1980) 'Graduation requirements in the state of Oregon: A case study', in R.M. JAEGER and C.K. TITTLE (Eds) *Minimum Competency Achievement Testing: Motives, Models, Measures and Consequences*, Berkeley, CA, McCutchan, pp. 258–63.

HILLARD, A.G. (1984) 'IQ testing as the emperor's new clothes: A critique of Jensen's *Bias in Mental Testing*', in C.R. REYNOLDS and R.T. BROWN (Eds) *Perspectives on Bias in Testing*, New York, Plenum, pp. 139–69.

HUNT, J. MCV. (1961) *Intelligence and Experience*, New York, Ronald Press.

HUNT, J. MCV. (1972) 'The role of experience in the development of competence', in J. MCV. HUNT (Ed.) *Human Intelligence*, New Brunswick, NJ, Transaction Books, pp. 30–52.

INTERNATIONAL READING ASSOCIATION (1979) 'A position on minimum competencies in reading', *Journal of Reading*, **23**, pp. 50–1.

JAEGER, R.M. (1987) 'Two decades of revolution in educational measurement!?' *Educational Measurement: Issues and Practice*, **6**, pp. 6–14.

JAEGER, R.M. and TITTLE, C.K. (Eds) (1980) *Minimum Competency Achievement Testing: Motives, Models, Measures and Consequences*, Berkeley, CA, McCutchan.

JENSEN, A.R. (1969) 'How much can we boost IQ and scholastic achievement?' *Harvard Educational Review*, **39**, pp. 1–123.

JENSEN, A.R. (1980) *Bias in Mental Testing*, New York, The Free Press.

KAMIN, L.J. (1974) *The Science and Politics of IQ*, Potomac, MD, Erlbaum.

KAUFMAN, A.S. and KAUFMAN, N.L. (1983) *Kaufman Assessment Battery for Children*, Circle Pines, MN, American Guidance Service.

KIRK, S.A. (1973) 'The education of intelligence', *Slow Learning Child*, **20**, pp. 67–83.

LAMBERT, N.M. (1981) 'Psychological evidence in *Larry P. v. Wilson Riles*: An evaluation by a witness for the defense', *American Psychologist*, **36**, pp. 937–52.

LERNER, B. (1981) 'The minimum competence testing movement: Social, scientific and legal implications', *American Psychologist*, **36**, pp. 1057–66.

LINN, R.L., MADAUS, G.F. and PEDULLA, J.J. (1982) 'Minimum competency testing: Cautions on the state of the art', *American Journal of Education*, **91**, pp. 1–35.

LYMAN, H.B. (1986) *Test Scores and What They Mean*, 4th ed., Englewood Cliffs, NJ, Prentice-Hall.

LYONS, C.H. (1975) *To Wash an Aethiop White: British Ideas about Black African Educability, 1530–1960*, New York, Teachers College Press.

MCCARTHY, D. (1972) *Manual for the McCarthy Scales of Children's Abilities*, New York, Psychological Corporation.

MCDILL, E.L., NATRIELLO, G. and PALLAS, A.M. (1985) 'Raising standards and retaining students: The impact of the reform recommendations on potential dropouts', *Review of Educational Research*, **55**, pp. 415–33.

MADAUS, G.F. (1986) 'Measurement specialists: Testing the faith — a reply to Mehrens', *Educational Measurement: Issues and Practices*, **5**, pp. 11–20.

MERCER, J.R. (1988) 'Ethnic differences in IQ scores: What do they mean?' (A response to Lloyd Dunn) *Hispanic Journal of Behavioral Sciences*, **10**, pp. 199–218.

MESSICK, S. (1989) 'Meaning and values in test validation: The science and ethics of assessment', *Educational Researcher*, **18**, pp. 5–11.

MIELE, F. (1979) 'Cultural bias in the WISC', *Intelligence*, **3**, pp. 149–54.

MILLIKEN, W.G. (1970) 'Making the school system accountable', *Compact*, **4**, pp. 17–18.

MITCHELL, J.V. (1984) 'Testing and the Oscar Buros lament: From knowledge to implementation to use', in B.S. PLAKE (Ed.) *Social and Technical Issues in Testing: Implications for Test Construction and Usage*, Hillsdale, NJ, Erlbaum, pp. 111–26.

NATIONAL COMMISSION ON EXCELLENCE IN EDUCATION (1983) *A Nation at Risk: The Imperatives for Educational Reform*, Washington, DC, US Government Printing Office.

NAVA, R. (1985) 'Caveat: Teacher competency tests may be hazardous to the employment of minority teachers and the education of language minority students', *Thrust*, **14**, pp. 33–4.

NITKO, A.J. (1984) 'Defining "criterion-referenced test"', in R.A. BERK (Ed.) *A Guide to Criterion-referenced Test Construction*, Baltimore, MD, Johns Hopkins University Press, pp. 8–28.

OAKES, J. (1985) *Keeping Track: How Schools Structure Inequality*, New Haven, CT, Yale University Press.

OAKLAND, T. and FEIGENBAUM, D. (1979) 'Multiple sources of test bias on the WISC-R and the Bender-Gestalt test', *Journal of Consulting and Clinical Psychology*, **47**, pp. 968–74.

OAKLAND, T. and GOLDWATER, D.L. (1979) 'Assessment and interventions for mildly retarded and learning disabled children', in G.D. PHYE and D.J. RESCHLY (Eds) *School Psychology: Perspectives and Issues*, New York, Academic Press, pp. 125–55.

OAKLAND, T. and LAOSA, L.M. (1977) 'Professional, legislative and judicial influences on psychoeducational assessment practices in schools', in T. OAKLAND (Ed.) *Psychological and Educational Assessment of Minority Children*, New York, Brunner/Mazel, pp. 21–51.

OLSEN, L. (1988) *Crossing the Schoolhouse Border: Immigrant Students and the California Public Schools*, Boston, MA, California Tomorrow.

ORFIELD, G. (1988) 'The Growth and Concentration of Hispanic Enrollment and the Future of American Education', paper presented at the National Council of La Raza Conference, Albuquerque, NM, July.

ORUM, L.S. (1986) *The Education of Hispanics: Status and Implications*, Washington, DC, National Council of La Raza.

PALLAS, A.M., NATRIELLO, G. and MCDILL, E.L. (1988) 'Who falls behind: Defining the 'at-risk' population — current dimensions and future trends', paper presented at the meeting of the American Educational Research Association, New Orleans, LA, April.

PAULSON, D. and BALL, D. (1984) 'Back to basics: Minimum competency testing and its impact on minorities', *Urban Education*, **19**, pp. 5–15.

PETERSEN, N.S. and NOVICK, M.R. (1976) 'An evaluation of some models for culture-fair selection', *Journal of Educational Measurement*, **13**, pp. 3–29.

PETERSON, P.L. (1982) 'Individual differences', in H.E. MITZEL (Ed.) *Encyclopedia of Educational Research*, 5th ed., New York, McMillan, pp. 844–51.

PHILLIPS, J. (1989) 'Students' test scores improve in AISD', *Austin American-Statesman*, August 28, pp. B1, B4.

PIAGET, J. (1952) *The Origins of Intelligence in Children*, New York, W.W. Norton.

RAFFERTY, J.R. (1988) 'Missing the mark: Intelligence testing in Los Angeles public schools', *History of Education Quarterly*, **28**, pp. 73–93.

RESCHLY, D.J. (1978) 'WISC-R factor structures among Anglos, Blacks, Chicanos, and Native American Papagos', *Journal of Consulting and Clinical Psychology*, **46**, pp. 417–22.

RESCHLY, D.J. (1979) 'Nonbiased assessment', in G.D. PHYE and D.J. RESCHLY (Eds) *School Psychology: Perspectives and Issues*, New York, Academic Press, pp. 215–53.

RESCHLY, D.J. (1980) 'Assessment of exceptional individuals: Legal mandates and professional standards', in R.K. MULLIKEN and M.R. EVANS (Eds) *Assessment of Children with Low-incidence Handicaps*, Washington, DC, National Association of School Psychologists, pp. 8–23.

RESCHLY, D.J. and RESCHLY, J.E. (1979) 'Validity of WISC-R factor scores in predicting achievement and attention for four sociocultural groups', *Journal of School Psychology*, **17**, pp. 355–61.

RESCHLY, D.J. and SABERS, D. (1979) 'Analysis of test bias in four groups with the regression definition', *Journal of Educational Measurement*, **16**, pp. 1–9.

RESNICK, D.P. (1981) 'Testing in America: A supportive environment', *Phi Delta Kappan*, **62**, pp. 625–8.

RESNICK, L.B. (1979) 'The future of IQ testing in education', *Intelligence*, **3**, pp. 241–53.

REYNOLDS, C.R. (1982) 'The problem of bias in psychological assessment', in C.R.

REYNOLDS and T.B. GUTKIN (Eds) *The Handbook of School Psychology*, New York, Wiley, pp. 178–208.

REYNOLDS, C.R. (1983) 'Test bias: In God we trust; all others must have data', *Journal of Special Education*, **17**, pp. 241–60.

REYNOLDS, C.R. and BROWN, R.T. (Eds) (1984) *Perspectives on Bias in Testing*, New York, Plenum.

REYNOLDS, C.R. and GUTKIN, T.B. (1980) 'A regression analysis of test bias on the WISC–R for Anglos and Chicanos referred to psychological services', *Journal of Abnormal Child Psychology*, **8**, pp. 237–43.

SALVIA, J. and YSSELDYKE, J.E. (1988) *Assessment in Special and Remedial Education*, 4th ed., Boston, MA, Houghton Mifflin.

SANDOVAL, J. (1979) 'The WISC–R and internal evidence of test bias with minority children', *Journal of Consulting and Clinical Psychology*, **47**, pp. 919–27.

SCHEUNEMAN, J.D. (1984) 'A theoretical framework for the exploration of causes and effects of bias in testing', *Educational Psychologist*, **19**, pp. 219–25.

SEBOW, R.C. (1984) 'Effects of minimum competency testing for minority students: A review of expectations and outcomes', *The Urban Review*, **16**, pp. 67–75.

SHEPARD, L.A. (1982) 'Definitions of bias', in R.A. BERK (Ed.) *Handbook of Methods for Detecting Test Bias*, Baltimore, MD, Johns Hopkins University Press, pp. 9–30.

SHEPARD, L.A. and KREITZER, A.E. (1987) 'The Texas teacher test', *Educational Researcher*, **16**, pp. 22–31.

SMITH, G.P. (1987) *The Effects of Competency Testing on the Supply of Minority Teachers*, a report prepared for the National Education Association and the Council of Chief State School Officers.

TERMAN, L.M. (1920) 'The use of intelligence tests in the grading of school children', *Journal of Educational Research*, **1**, pp. 20–32.

THORNDIKE, R.L. (1971) 'Concepts of cultural fairness', *Journal of Educational Measurement*, **8**, pp. 63–70.

THORNDIKE, R.L. and HAGEN, E. (1977) *Measurement and Evaluation in Psychology and Education*, New York, Wiley.

TITTLE, C.K. (1982) 'Use of judgmental methods in item bias studies', in R.A. BERK (Ed.) *Handbook of Methods for Detecting Test Bias*, Baltimore, MD, Johns Hopkins University Press, pp. 31–63.

TYACK, D. (1974) *The One Best System: A History of American Urban Education*, Cambridge, MA, Harvard University Press.

ULICH, R. (1950) *History of Educational Thought*, New York, American Book Co.

US COMMISSION ON CIVIL RIGHTS (1974) *Mexican American Education Study, Report 6: Toward Quality Education for Mexican Americans*, Washington, DC, Government Printing Office.

VALENCIA, R.R. (1982) 'Psychoeducational needs of minority children: The Mexican American child, a case in point', in S. HILL and B.J. BARNES (Eds) *Young Children and Their Families: Needs of the 80s*, Lexington, MA, Lexington Books, D.C. Heath, pp. 73–87.

VALENCIA, R.R. (1984) 'Concurrent validity of the Kaufman Assessment Battery for Children in a sample of Mexican American children', *Educational and Psychological Measurement*, **44**, pp. 365–72.

VALENCIA, R.R. (1985a) 'Stability of the Kaufman Assessment Battery for Children in a sample of Mexican American Children', *Journal of School Psychology*, **23**, pp. 189–93.

VALENCIA, R.R. (1985b) 'Predicting academic achievement with the Kaufman Assessment Battery for Children in Mexican American children', *Educational and Psychological Research*, **5**, pp. 11–17.

VALENCIA, R.R. (1988) 'The McCarthy Scales and Hispanic children: A review of psychometric research', *Hispanic Journal of Behavioral Sciences*, **10**, pp. 81–104.

VALENCIA, R.R. (1990) *Chicano Intellectual Performance: Theory, Research, and Schooling Implications*, manuscript in progress.

VALENCIA, R.R. and ABURTO, S. (in press, a) 'Competency testing and Latino student access to the teaching profession: An overview of issues', in J. DENEEN, G.D. KELLER and R. MAGALLAN (Eds) *Assessment and Access: Hispanics in Higher Education*, Albany, NY, State University of New York Press.

VALENCIA, R.R. and ABURTO, S. (in press, b) 'Research directions and practical strategies in teacher testing and assessment: Implications for improving Latino access to teaching', in J. DENEEN, G.D. KELLER and R. MAGALLAN (Eds) *Assessment and Access: Hispanics in Higher Education* Albany, NY, State University of New York Press.

VALENCIA, R.R. and RANKIN, R.J. (1985) 'Evidence of content bias on the McCarthy Scales with Mexican American children: Implications for test translation and nonbiased assessment', *Journal of Educational Psychology*, **77**, pp. 197–207.

VALENCIA, R.R. and RANKIN, R.J. (1986) 'Factor analysis of the K-ABC for groups of Anglo and Mexican American children', *Journal of Educational Measurement*, **23**, pp. 209–19.

VALENCIA, R.R. and RANKIN, R.J. (1988) 'Evidence of bias in predictive validity on the Kaufman Assessment Battery for Children in samples of Anglo and Mexican American children', *Psychology in the Schools*, **22**, pp. 257–63.

VALENCIA, R.R. and RANKIN, R.J. (1990) 'Examination of Content Bias on the K-ABC with Anglo and Mexican American Children', manuscript submitted for publication.

WATSON, A. (1989) 'Minority teachers sought', *San Jose Mercury News*, December 31, pp. 1A, 6A.

WATSON, A. and KRAMER, P. (1989) 'CAP math scores up, reading down in state: Honig pleased to see overall trend rising', *San Jose Mercury News*, April 26, pp. 1B–2B.

WECHSLER, D. (1974) *Manual for the Wechsler Intelligence Scale for Children — Revised*, New York, Psychological Corporation.

WEINTRAUB, F.J. and ABESON, A.R. (1972) 'Appropriate education for all handicapped children: A growing issue', *Syracuse Law Review*, **23**, pp. 1037–58.

WIERSMA, W. and JURS, S.G. (1985) *Educational Measurement and Testing*, Boston, MA, Allyn and Bacon.

WILLIAMS, R. (1971) 'Abuses and misuses in testing black children', *The Counseling Psychologist*, **2**, pp. 62–73.

ZAPATA, J.T. (1988) 'Early identification and recruitment of Hispanic teacher candidates', *Journal of Teacher Education*, **39**, pp. 19–23.